# Gestion de données personnelles respectueuse de la vie privée

Nicolas Anciaux

# Gestion de données personnelles respectueuse de la vie privée

**Nicolas ANCIAUX**

**Rapport scientifique pour l'obtention de l'**

**HABILITATION A DIRIGER LES RECHERCHES EN INFORMATIQUE**

**Université de Versailles Saint-Quentin-en-Yvelines**

**Soutenance prévue le 9 décembre 2014**

---

**Jury**

| | | |
|---|---|---|
| **Rapporteurs :** | **Serge ABITEBOUL** | **(Directeur de Recherche, Inria, & Prof., ENS Cachan)** |
| | **Ernesto DAMIANI** | **(Prof., Université de Milan)** |
| | **David GROSS-AMBLARD** | **(Prof., Université de Rennes 1)** |
| **Membres :** | **Philippe AIGRAIN** | **(Dr. HDR, La Quadrature du Net, CEO Sopinspace)** |
| | **Philippe PUCHERAL** | **(Prof., U. de Versailles St-Quentin-en-Yvelines)** |
| **Tuteur :** | **Luc BOUGANIM** | **(Directeur de Recherche, Inria)** |

« *(…) il ne faut pas faire de l'histoire pour l'histoire, on peut le faire, mais c'est comme faire de la philosophie pour la philosophie, de la musique pour la musique ou de faire de l'art pour l'art, c'est s'installer dans une position d'esthète; il faut faire de l'histoire pour faire de telle sorte que notre présent et notre futur soient possibles autrement que ce qu'a été le passé, bien sûr, ça n'aurait aucun sens de faire l'archéologie du passé, si ça ne permettait pas l'architecture du futur, voire l'architecture du présent* ».

**Michel Onfray**

Contre histoire de la philosophie, Saison 12 : Pensée post-nazie

Le principe Eichman de notre monde

Analyse d'un petit texte de Gunter Anders: "Nos fils d'Eichman"

Podcast: http://www.franceculture.fr/player/reecouter?play=4870664

La citation se trouve à 54 minutes 28 secondes

Entendu à la radio, mercredi 28 août à 20h55, soir où j'ai terminé l'écriture de mon manuscrit, et qui m'est apparue pertinente pour certains de nos travaux de recherche en informatique.

# Préliminaire

La majorité des résultats présentés dans ce document sont le fruit de collaborations, comme en atteste les publications sur lesquelles ils reposent. Je voudrais donc exprimer toute ma reconnaissance à mes co-auteurs et à tous ceux et toutes celles qui ont contribué à ces travaux.

# Table des matières

# Introduction

*Ce manuscrit présente certains de mes résultats de recherche, relatifs à l'élaboration d'un nouveau modèle de gestion des données personnelles, plus respectueux de la vie privée. Les travaux présentés sont volontairement introduits de manière informelle, espérant les rendre accessibles à un non spécialiste. Ce chapitre d'introduction présente la motivation de ces travaux, les axes de recherche considérés, résume mon parcours scientifique, les projets dans lesquels je m'implique, et introduit le plan du document.*

## 1. Motivation des travaux de recherche

En très peu de temps, nous sommes entrés dans une ère de génération massive des données personnelles créées par les individus, leurs équipements digitaux (smartphones, équipements d'auto mesure, compteurs électriques intelligents) ou mises à disposition par les organisations (banques, administrations, centres médicaux, etc.). L'ensemble de ces données constitue la vie numérique (souvent privée) de l'individu, décrivant ses déplacements, sa consommation, ses relations, son état médical, social, financier, ses comportements, ses préoccupations, etc.

Ces données constituent une manne pour l'économie. La valeur boursière des entreprises dont le modèle d'affaire est basé sur l'exploitation des données personnelles en témoigne. Deux milliards de dollars par an sont dépensés aux Etats-Unis dans l'achat d'informations personnelles [Kha11]. Le Forum Economique Mondial assimile les données personnelles à un « nouveau pétrole » [WEF11] et certaines initiatives politiques les assimilent à une ressource et envisagent des manières de les taxer pour qu'elles n'échappent pas à la TVA [CoC13].

Les individus, pour pouvoir bénéficier de leurs propres données, passent par des applications en ligne qui rendent ces données exploitables et accessibles. Le respect de la vie privée des individus est alors délégué à l'application, de manière contractuelle, mais sans garantie tangible pour l'individu. Les données personnelles sont ainsi parfois exploitées de façon très peu transparentes à des fins secondaires, pour répondre aux exigences des modèles d'affaire ayant cours. Des passe-droits peuvent être accordés pour satisfaire les requêtes d'un gouvernement ou d'un partenaire industriel. Et tous ces usages peuvent être menés de façon unilatérale par le gestionnaire des données, sans l'assentiment de l'individu concerné, ou suivant des chartes de confidentialité parfois floues, changeantes, et présumées acceptées par celui-ci. Les systèmes

manquent souvent aussi d'ouverture pour l'usager vis-à-vis de ses propres données, et tout désengagement est souvent difficile ou pratiqué au risque de perdre ses données. D'autre part, la centralisation des données personnelles conduit à des divulgations accidentelles et à des attaques informatiques répétées impactant de très grands volumes de données. Ainsi, l'utilisateur, dépossédé de tout moyen de contrôle, ne peut ni éviter ni même souvent connaître les usages indésirables qui pourraient être faits de ses propres données.

La situation actuelle est donc très contestable du point de vue du respect de la vie privée et cela commence à impacter l'économie. La révélation de l'observation du Cloud par la NSA pourrait conduire à des pertes économiques pour les acteurs du Cloud Américain estimées entre 22 et 180 milliards de dollars selon les analystes [Cas13, Sta13]. IBM implante des serveurs hors des Etats-Unis [Mil14], pour satisfaire ses clients pour qui la localisation géographique (et les passe-droits associés) prend de l'importance. Le Forum Economique Mondial lui-même plaide pour un contrôle accru des usagers sur leurs données [WEF12].

Un consensus économique, politique et social émerge actuellement, pour parvenir à un modèle plus respectueux de la vie privée. Les grands acteurs économiques travaillent dans cet objectif comme en attestent les travaux du groupe « Trustworthy Computing » de Microsoft[1] visant à améliorer la confiance dans les serveurs. De plus, certains industriels sont réticents à une exploitation généralisée des données personnelles qu'ils ont en charge dans le modèle actuel du Web et soutiennent les approches visant à tirer parti de l'explosion actuelle de l'utilisation des données personnelles digitales, tout en préservant l'intimité des usagers. Par exemple, EDF qui se voit comme un tiers de confiance pour les données de consommation électrique, intègre actuellement des technologies de protection de la vie privée dans ses compteurs électriques intelligents. Nous assistons aussi à une prise de conscience du monde politique en Europe qui s'exprime notamment par le biais du renforcement[2] de la Directive 95/46/CE [Res14, Dir95] qui édicte les principes légaux de respect de la vie privée numérique dans l'Union. Une décision de

[1] Voir http://www.microsoft.com/en-us/twc/default.aspx

[2] CNIL. Règlement européen et surveillance des citoyens : avancées au Parlement européen. http://www.cnil.fr/linstitution/actualite/article/article/reglement-europeen-et-surveillance-des-citoyens-avancees-au-parlement-europeen/

la Cours Européenne de mars dernier[3], contraignant Google à offrir aux usagers un outil de droit à l'oubli, va aussi dans le sens d'un durcissement. Enfin, les représentants de la société civile et de nombreux usagers restent attachés aux principes de respect de la vie privée numérique. Contrairement à l'idée reçue, les plus jeunes ne sont pas moins préoccupés que leurs aînés par le respect de leur intimité [BBD14] et sont plus nombreux aujourd'hui à modifier les paramètres de confidentialité de leur smartphone[4] et à bloquer des applications perçues comme trop invasives [MLC+13]. De plus, ils se tournent vers des moyens de communication plus éphémères comme Snapchat qui permet d'envoyer des photos qui disparaissent quelques secondes après avoir été visualisées.

Ainsi, les données personnelles doivent être manipulées sous un contrôle accru des individus de manière à rétablir la confiance nécessaire. Il s'agit donc de garantir les principes fondamentaux du respect de la vie privée : consentement des individus pour la finalité du traitement, collecte et rétention de données limitées, exposition minimale des données à des tiers, droit d'ouverture sur les données et audit des usages.

Mais il n'y a pas encore de solution technique satisfaisante. Les deux approches actuelles consistent à améliorer la confiance que l'on peut mettre dans les serveurs ou à introduire des serveurs personnels en charge de la gestion des données de leur propriétaire. La première approche ne permet pas de résoudre les problèmes intrinsèques aux approches centralisées (attaques sophistiquées, modèle basé sur la délégation) et les approches décentralisées sacrifient les fonctionnalités et usages innovants sans toutefois garantir une très grande sécurité.

Nous introduisons une nouvelle approche que nous appelons le « Web Personnel Sécurisé » où les individus régulent leurs données personnelles depuis des composants personnels sécurisés. Les fonctionnalités principales des solutions centralisées doivent être préservées: durabilité, disponibilité, partage des données. Mais l'exploitation des données se fait avec

---

[3] Le Monde.fr, 13 mai 2014, « Droit à l'oubli : Google débouté par la justice européenne », par Martin Untersinger. http://www.lemonde.fr/technologies/article/2014/05/13/droit-a-l-oubli-google-deboute-par-la-justice-europeenne_4415804_651865.html

[4] Snowden effect: Young people now care about privacy. By Byron Acohido, USA Today, 18 Nov. 2013.

l'assentiment du propriétaire, qui régule les usages au travers des autorisations qu'il donne, et dispose de fortes garanties de non contournement de ses directives.

## 2. Travaux de recherche

Mes travaux de recherche s'intègrent dans les deux axes de recherche sous-jacents au modèle du Web personnel sécurisé. Il s'agit d'une part (***Axe 1***), d'embarquer des techniques de gestion de données dans les composants personnels sécurisés pour en faire de véritables serveurs personnels de données, et d'autre part (***Axe 2***) de définir et mettre en œuvre de nouveaux modèles de gestion de données respectueux de la vie privée, régulant le partage, la collecte, la rétention et l'usage des données personnelles, et d'intégrer ces modèles dans une architecture suivant une approche Privacy-by-Design offrant de fortes garanties de non contournement à l'utilisateur.

***Axe 1***. Concernant la gestion de données embarquées, nous nous concentrons sur des composants sécurisés matériellement, notamment sur des dispositifs dotés de microcontrôleurs sécurisés et disposant d'une mémoire Flash (grande capacité de stockage). Le problème est de concevoir des algorithmes de gestion de données et des structures accélératrices, de façon adaptée aux fortes contraintes du composant: très faible quantité de RAM ; caractéristiques techniques des mémoires Flash induisant des performances particulières d'accès en lecture/écriture. Mes contributions sur le sujet portent sur la conception d'un système de gestion de bases de données (SGBD) relationnel embarqué [ABP+14, ABG+10, AAB+07, SAB+07, AAB+09].

***Axe 2***. Concernant la définition de procédés de gestion de données respectueux de la vie privée, certains de mes travaux ont porté sur la gestion de données dans une base de données intégrant des limites de rétention suivant les objectifs des traitements [HFA09, ABH+08a, ABH+08b, ABH+08c, HAF+06] et ont été conduits dans le cadre d'une coopération avec l'Université de Twente (Pays-Bas). Les contributions les plus récentes se concentrent sur la définition d'architectures Privacy-by-Design reposant sur des puces sécurisées [ABD+14a, ABD+14b, ABB+13, ANP13, AAB+10a], sur des procédés d'exposition minimum lors d'interactions avec des processus externes de prise de décision personnalisée [ABN+15, ABN+13, ABN+12, ANV12] et sur de nouveaux modèles permettant aux individus de réguler le partage de leurs données [AAB+10b, AAB+09, AAB+10c].

Ces deux axes de recherche sont très complémentaires : la gestion ubiquitaire de données personnelles introduit de nouveaux problèmes de préservation de la vie privée, et la gestion de données embarquées dans des composants sécurisés permet d'envisager de nouveaux modèles de sécurisation de bases de données.

Outre les problèmes techniques, ce thème de recherche touche de nombreux enjeux économiques, juridiques, ou sociologiques. Nous ne prétendons pas étudier ces enjeux, mais nous essayons de confronter nos solutions techniques à ces enjeux. Cela passe par une stratégie de validation de nos propositions, des démonstrations aux industriels, des discussions avec des chercheurs d'autres disciplines (droit, économie, etc.) et des essais sur le terrain impliquant des usagers. Ainsi, le prototypage, l'expérimentation, et les discussions multi disciplinaires font partie intégrante de mon activité de recherche.

## 3. Parcours

J'obtiens ma thèse de doctorat [Anc04] de l'Université de Versailles Saint-Quentin en Yvelines fin 2004 sous la direction de Philippe Pucheral, Professeur à l'Université de Versailles Saint-Quentin-en-Yvelines. J'aborde dans ma thèse l'étude de l'environnement carte à puce sous l'angle de la gestion de données, l'évaluation de requête en environnement contraint, les techniques de co-design permettant de calibrer à la fois les ressources matérielles de la puce (notamment la RAM) à une application donnée, et inversement, les structures de données et les opérateurs internes à la quantité de RAM disponible [ABP+01, ABP03b] pour parvenir à la définition de bancs d'essais pour SGBD embarqués [ABP+08b]. Des contacts avec Bull CP8 puis Axalto (maintenant intégré à Gemalto) me permettent alors d'avoir accès à des prototypes de carte à puce avancés dans lesquels j'ai pu porter mon implémentation de PicoDBMS [ABP+01], le premier SGBD complet (supportant l'algèbre relationnelle) tournant sur carte à puce. Ce code a également permis à Axalto de tester son système d'exploitation embarqué et l'a conduit à appliquer certaines optimisations.

J'obtiens en 2005 un poste de chercheur post-doctorant dans l'équipe bases de données du département d'informatique et du CTIT (Center for Telematics and Information Technology) de l'Université de Twente dirigé par Peter Apers, Professeur à l'Université de Twente. Ma

recherche s'inscrit alors dans le cadre du projet national NWO-VIDI-2005 intitulé "*Context-Aware Data Management Towards Ambient Intelligence*" conduit par Ling Feng, maintenant Professeur à l'Université Tsinghua en Chine. Mon rôle était d'étudier la préservation de l'intimité des usagers évoluant dans un environnement d'intelligence ambiante doté de nombreux dispositifs communicants. Plus particulièrement, j'ai proposé des procédés de dégradation progressive et automatique des données personnelles, fondés sur l'hypothèse que les objectifs à long terme des applications peuvent être atteints en utilisant des données plus générales (moins précises) que leurs objectifs à court terme. J'ai encadré le stage d'Harold van Heerde sur cette thématique puis sa thèse (de 2006 à 2010) en collaboration avec Maarten Fokkinga, enseignant chercheur à l'Université de Twente, sous la co-direction de Peter Apers et de Philippe Pucheral. Nos travaux menés dans le cadre de cette thèse ont donné lieu aux publications [HFA09, ABH+08a, ABH+08b, ABH+08c, HAF+06].

Ayant pris pleinement conscience de l'importance sociétale croissante des problèmes liés au respect de la vie privée et souhaitant me consacrer pleinement à leur étude, je postule au concours de Chargé de Recherche INRIA et suis recruté en 2006 dans le projet SMIS[5]. A mon arrivée nous montons les projets PlugDB et CG78/DMSP et j'encadre les travaux menés par les ingénieurs et doctorants dans le cadre de ces projets. Nous abordons alors le problème de la gestion sécurisée de données embarquées et cherchons à concevoir un serveur de données sécurisé. Nous nous intéressons d'abord avec Dennis Shasha, Professeur à l'Université de New York, au problème des traitements distribués entre une base de données embarquée et confidentielle (statique, en lecture seule) et une base de données externe et publique [AAB+07, SAB+07, AAB+09]. Nous démarrons les thèses de Yanli Guo (2008-2011) et de Lionel le Folgoc (2009-2012), que je co-encadre sous la direction de Luc Bouganim, sur la conception d'un moteur de gestion de données embarqué dans un microcontrôleur sécurisé relié à une mémoire Flash externe de grande capacité (Go). Nous jetons ensuite les bases d'une première architecture de gestion de données personnelles respectueuse de la vie privée, conçue dans une approche Privacy-by-Design, en collaboration avec Indrajit et Indrakshi Ray, Professeurs à l'Université de Colorado. Cette étude bénéficie de l'expérience développée dans le cadre des projets PlugDB et CG78/DMSP qui nous sert de cas d'usage. Les travaux sur le Serveur

[5] Secured and Mobile Information Systems, équipe commune INRIA-UVSQ-CNRS. http://www-smis.inria.fr/

Personnel de Données embarqué donnent lieu aux publications [ABP+14, ABG+10], la première version d'une architecture Privacy-by-Design est présentée à la conférence VLDB'10 [AAB+10a]. Nous appliquons aussi ces travaux au cas des données de santé [AAB+10b, AAB+10c, ABB+08a, ABB+08b]. Nous cherchons aussi à étendre nos techniques de gestion de données embarquées, pensées au départ pour le modèle relationnel, à d'autres modèles de données. En 2012, nous lançons la thèse de Saliha Lallali sur cette thématique. Nous cherchons à généraliser les techniques de gestion de données embarquées pour couvrir au moins le cas de l'indexation de documents.

Nous nous posons aussi la question du contrôle des données hors du serveur personnel. En 2011, nous démarrons avec Michalis Vazirgiannis, Professeur à l'Ecole Polytechnique, une étude sur le problème de l'exposition du minimum de données personnelles lorsqu'un serveur personnel interagit avec l'extérieur. Ces travaux donnent lieu aux publications [ABN+15, ABN+13, ABN+12, ANV12]. Nous lançons en parallèle une autre étude, en collaboration avec Philippe Bonnet, Professeur à l'ITU au Danemark, sur le contrôle d'usage des données manipulées en dehors du serveur personnel. Une architecture préliminaire a été présentée à CIDR'13 [ABB+13]. Depuis fin 2013 nous envisageons un usage du serveur personnel adapté au contexte des Pays les Moins Avancés, pouvant fonctionner sans aucune infrastructure (réseau, PKI, etc.). J'ai présenté notre vision de ce type d'usage à VLDB'14 [ABD+14a]. Nous étudions actuellement avec le LIRIMA (laboratoire Inria en Afrique) la possibilité de travailler conjointement sur cette base.

## 4. Implication dans des projets

Les projets auxquels je participe sont tous positionnés sur des thématiques liées à la gestion de données respectueuse de la vie privée. Certains de ces projets sont conduits avec des partenaires d'autres disciplines (juristes, économistes, et sciences humaines et sociales), des partenaires industriels et des représentants de la société civile, pour nous permettre de bien appréhender les aspects transverses liés à la dimension sociétale de notre thématique.

**Projet CG78/DMSP (Département des Yvelines, depuis 2007)**

https://project.inria.fr/plugdb/

***Coordinateur***: Philippe Pucheral et Nicolas Anciaux (SMIS).

***Mon rôle***: coordination du projet, référent technique du projet.

***Partenaires***: Inria, Université de Versailles, Santeos (Atos Origin), Gemalto, ALDS (coordination gérontologique médicale), et COGITEY (coordination sociale).

***Objectif***: Le projet a pour objectif de concevoir un dossier médico-social mobile et sécurisé facilitant les soins au domicile de personnes dépendantes, et d'expérimenter la solution sur le territoire des Yvelines. Le projet a déjà donné lieu à 3 conventions : 2007-2010 (élaboration de la technologie), 2011-2012 (expérimentation terrain) et 2013-2014 (évolution de la technologie). Au niveau technique, le projet implique la conception et la mise en œuvre d'un serveur personnel de données sur un composant matériel combinant un microcontrôleur sécurisé (type carte à puce) et une grande quantité de la mémoire FLASH dans un format carte SIM, ainsi que la conception et le développement des services attenants de synchronisation et de restauration du serveur embarqué. L'expérimentation terrain s'est déroulée sur 18 mois en 2011-2012 auprès d'une centaine de patients et professionnels médicaux sociaux sur le territoire des Yvelines. Un retour d'expérience a conduit à des adaptations importantes (matérielles et logicielles) du composant personnel sécurisé nous amenant à faire fabriquer nous-mêmes un nouveau composant doté d'une interface Bluetooth et d'un lecteur d'empreinte digitale. L'ARS Ile de France réalise actuellement un audit de la solution développée dans le projet afin d'envisager un déploiement plus large (résultat de l'audit prévu pour fin 2014). Une vidéo décrit la solution et une démonstration est disponible. La version actuelle du composant personnel s'interface avec n'importe quel Smartphone ou tablette Android équipé d'un port USB ou du Bluetooth.

### Projet CityLab@Inria (Inria Project Lab, depuis juin 2014)

https://citylab.inria.fr/

***Coordinateur***: Valérie Issarny (Inria@Sillicon Valley & Arles-Mimove).

***Mon rôle*** : Responsable pour le partenaire SMIS.

***Partenaires***: Arles-Mimove, Clime, Dice, Fun, Myriads, OAK, SMIS, Urbanet et Willow.

***Objectif***: Le projet étudie les solutions ICT pour la ville intelligente dans un objectif de soutenabilité sociale (et environnementale). J'ai participé à la rédaction de la proposition de projet et y représente l'équipe SMIS. Notre implication a pour but d'envisager des architectures Privacy-by-Design garantissant la vie privée des citoyens dans un contexte où

ils sont producteurs de données [ABB+14]. Nous nous intéressons en particulier à la capture de données sociales, produites par les usagers depuis leur smartphone, dans un environnement dit de "social sensing".

**ISN (Idex Paris Saclay, depuis dec. 2013)**

http://digitalsocietyinstitute.com/

***Coordinateurs du pôle***: Fabrice Le Guel (ADIS) et Benjamin Nguyen (SMIS).

***Mon rôle*** : Membre du pôle « Vie privée et identité numérique » et responsable pour SMIS du projet PEPS PAIP.

***Partenaires***: GRACE/LIX, COMETE/LIX, DANTE, CERDI, SAMOVAR, SMIS, RITM.

***Objectif***: L'Institut de la Société Numérique (ISN) adopte une approche interdisciplinaire, entre disciplines informatiques et sciences humaines économiques et sociales, pour étudier certains défis sociétaux inhérents à la société numérique. Deux pôles ont été lancés: le premier sur le thème de la co-évolution homme/machine, le second sur celui de la vie privée et l'identité numérique dans lequel SMIS est impliqué. Nous avons notamment lancé un projet PEPS financé par le CNRS impliquant les partenaires du pôle dans lequel nous évaluons, sous forme expérimentale, l'impact sur les usagers de solutions de gestion de données personnelles où l'individu possède (physiquement) le serveur qui régit la dissémination de ses données, par rapport aux solutions centralisées classiques.

**Projet KISS (ANR INS, Dec. 2011 – Dec. 2015)**

https://project.inria.fr/kiss/

***Coordinateur***: Philippe Pucheral (SMIS).

***Mon rôle*** : Responsable de la tâche sur l'exposition minimum de données.

***Partenaires***: Conseil Général des Yvelines, CryptoExpert, Gemalto, Inria (SMIS & SECRET), LIRIS, PRiSM.

***Objectif***: Le projet vise à produire une alternative crédible à la centralisation systématique des données personnelles sur des serveurs tiers, ouvrant la voie à de nouvelles solutions suivant une approche Privacy-by-Design pour la gestion des données personnelles. L'idée soutenue dans KISS est d'embarquer dans des composants personnels de confiance, des modules logiciels capables d'acquérir, de stocker et de gérer différentes formes

d'informations personnelles (ex. bulletins de salaires, factures, données bancaires, médicales, traces de géolocalisation) selon les applications, et d'en réguler la dissémination [APP+12]. Ces modules logiciels forment un serveur personnel de données capable de s'interfacer avec des services externes mais restant sous le contrôle de son propriétaire. Je suis responsable dans ce projet des travaux cherchant à limiter au minimum les données à exposer à des services externes. Nous avons proposé des procédés et algorithmes adaptés à certains scénarios applicatifs validés avec le Conseil Général des Yvelines dans le cadre de la demande d'aide sociale.

**Projet DEMOTIS (ANR-ARPEGE, 2009 – 2012)**

http://www.demotis.org/

***Coordinateur***: Philippe Aigrain (Sopinspace).

***Mon rôle*** : Responsable scientifique pour les équipes Inria.

***Partenaires*** : CECOJI, Inria (CACAO, SECRET, SMIS), Sopinspace.

***Objectif*** : Le projet DEMOTIS (Définir, Évaluer et MOdéliser les Technologies de l'Information de Santé) vise à éclairer les limitations et compromis réciproques que l'intrication des domaines juridiques et informatiques impose à la conception d'infrastructures en charge du Dossier Médical Personnalisé (DMP) et celles des dossiers des réseaux de soins liés à certaines affections (SIDA, cancer). Les deux volets du projet, juridique (droit de la santé, des données personnelles ou de la propriété intellectuelle) et informatique (sécurité des bases de données, techniques cryptographiques utilisées pour les protéger, ou anonymisation de données) ont été abordés de manière conjointe par les partenaires.

**Projet PlugDB (ANR RNTL, 2007 – 2010)**

***Coordinateur***: Philippe Pucheral (SMIS).

***Mon rôle*** : Responsable de la coordination technique.

***Partenaires*** : ALDS (coordination gérontologique médicale), Gemalto, Inria SMIS, PRiSM, Santeos (filiale d'Atos).

***Objectif*** : Conception d'un serveur personnel de données sur un nouveau composant matériel combinant un microcontrôleur sécurisé (type carte à puce) et une grande quantité de la

mémoire FLASH (Go) dans un châssis USB. La solution doit offrir une alternative à la centralisation des données plus respectueuse de la vie privée, tout en restaurant les propriétés classiques d'un serveur central.

**Projet CADMAI (NWO VIDI, 2005 – 2010)**

Grant individuel attribué au Professeur Ling Feng (Université de Twente, Pays-Bas)

***Mon rôle*** : Responsable de la tâche relative à la protection de la vie privée.

***Objectif*** : Le projet CADMAI (Context-Aware Data Management Towards Ambient Intelligence) étudie les problèmes liés à la conception et à la mise en œuvre de techniques de gestion de données dans un contexte d'intelligence ambiante. Ling Feng a imaginé le projet et formé une équipe de cinq doctorants et d'un chercheur post-doctorant. C'est dans le cadre de ce projet que j'ai réalisé mon post-doctorat. Je me suis principalement intéressé à l'intimité que peuvent avoir les individus dans un tel environnement et ai initié une étude sur la dégradation progressive des données personnelles. J'ai encadré en 2005/2006 le stage de Master d'Harold van Heerde sur la thématique qu'il a ensuite poursuivie en thèse.

## 5. Plan du document

Ce manuscrit présente un sous-ensemble de mes travaux de recherche. Il est organisé en trois volets liés à l'étude du modèle du Web Personnel Sécurisé présenté en introduction: *architecture* (chapitre 1), *contributions techniques* sous-jacentes (chapitres 2 et 3), et *application* (chapitre 4). Les deux premiers volets reposent sur les articles scientifiques annexés au document et le troisième décrit une application emblématique de ce que les techniques présentées dans ce manuscrit peuvent apporter et montre la faisabilité de l'approche. Le manuscrit est conçu comme un guide de lecture accessible aux non spécialistes présentant de façon informelle les travaux scientifiques placés en annexe. Le contenu de chacun des chapitres est résumé ci-dessous.

***Chapitre 1*** : ***Architecture.*** Nous introduisons dans ce chapitre une famille d'architectures radicalement différente de celle du Web actuel où l'individu exerce un contrôle sur ses données personnelles depuis des composants personnels sécurisés situés aux extrémités du réseau, tout en continuant à bénéficier des mêmes fonctionnalités qu'avec une solution centralisée. Le

chapitre introduit trois architectures représentatives de différents contextes. Chacune génère des problèmes scientifiques pour la communauté base de données dont certains sont étudiés dans les chapitres 2 et 3. Chacune des architectures présentée ici repose sur une publication annexée au document : l'architecture « Serveur Personnel de données » décrite dans la section 3.1 repose sur [AAB+10a] (annexe A), l'architecture à base de « Cellules de Confiance » décrite en section 3.2 repose sur [ABB+13] (annexe B), et l'architecture « Folk-IS » décrite en section 3.3 repose sur [ABD+14a] (annexe C).

***Chapitre 2*** : ***Serveur Personnel de Données.*** Ce chapitre se concentre sur la conception du Serveur Personnel de Données (SPD) embarqué formant le cœur des architectures présentées au chapitre 1. Nous considérons un dispositif qui, à l'image de nouveaux objets personnels qui fleurissent actuellement (carte SIM à grande capacité, capteur grande mémoire, carte SD sécurisée, etc.), combine un microcontrôleur sécurisé (type carte à puce) et une mémoire Flash de grande capacité (Go). Nous présentons dans ce chapitre les contraintes posées par cet environnement, leur impact sur la conception d'un SGBD relationnel embarqué dans le dispositif et les solutions que nous proposons. Ce chapitre repose sur la publication [ABP+14] (annexe D).

***Chapitre 3*** : ***Exposition Minimum.*** Ce chapitre montre comment l'introduction d'un SPD permet à un usager de réduire la dissémination de ses données au minimum lorsqu'il interagit avec un service externe. Nous nous plaçons dans le cas de la collecte de données via des formulaires telle qu'elle est pratiquée par les organisations (aide sociale, banques, etc.) souhaitant ajuster leur offre à la situation spécifique d'un demandeur. Nous montrons en quoi l'approche actuelle qui détermine a priori les données à divulguer, n'est pas minimale. Nous décrivons notre solution, basée sur une modélisation des objectifs du demandeur sous forme de règles de collecte, confrontées aux données du demandeur dans son SPD, pour permettre cette minimisation. Ce chapitre repose sur la publication [ANV12a] (annexe E).

***Chapitre 4*** : ***Application DMSP.*** Plusieurs applications respectueuses de la vie privée ont été développées par l'équipe SMIS. Cette section se focalise sur l'une d'entre elles, fondatrice pour l'équipe et particulièrement représentative de notre approche: l'application « Dossier Médico-Social Personnel » (DMSP). Elle repose sur une architecture très proche de l'architecture

« Serveur de Données Personnel » présentée section 3.1, chapitre 1. L'application a été développée pour le Conseil Général des Yvelines et a donné lieu à une expérimentation terrain. Ce chapitre présente, au travers de cette application, la faisabilité de notre approche, l'intérêt du modèle de gestion de données que nous proposons et les résultats principaux direct et indirects liés à cette application.

# Chapitre 1

# Architecture

*Notre approche consiste à introduire un serveur personnel sécurisé matériellement, permettant à l'individu d'exercer un contrôle sur ses données tout en préservant les avantages des solutions centralisées : durabilité, disponibilité et partage des données. Ce chapitre motive nos contributions architecturales, introduit trois propositions reposant sur les publications VLDB'10 [AAB+10a] présentée en annexe A, CIDR'13 [ABB+13] en annexe B, et VLDB'14 [ABD+14a] en annexe C et conclut en résumant les problèmes scientifiques sous-jacents abordés dans les chapitres suivants et nos résultats les plus significatifs.*

## 1. Motivation et état de l'art

Comme cela a été décrit en introduction, un large consensus existe aujourd'hui sur la nécessité de renforcer le contrôle des individus sur la gestion de leurs données personnelles, très insuffisant dans le modèle actuel du Web.

Deux approches principales sont considérées actuellement. La première, suivie par la plupart des grands éditeurs de systèmes de gestion de bases de données (IBM, Microsoft, Oracle, etc.), consiste à améliorer la confiance que les usagers placent dans le système en implantant dans les serveurs de nouvelles mesures renforçant le respect de la vie privée. IBM propose le concept de SGBD hippocratiques [AKS+02], rendant le serveur de données responsable de l'application des principes législatifs relatifs à la gestion de données personnelles (consentement, finalité, exposition limitée, collecte et rétention minimum, audit, etc.). Microsoft propose d'introduire le concept de serveurs de confiance (« Trustworhty Computing ») pour promouvoir l'implantation de mesures de sécurité accrues sur les serveurs [Cha12] : sécurisation matérielle apportée par les modules TPM, réduction du nombre de personnes ayant des droits administrateurs et implantation de principes attenants au respect de la vie privée et de structures de contrôle assermentant les systèmes. De plus, la plupart des grands éditeurs de bases de données ont intégré ces dernières années de nouvelles fonctionnalités liées à la sécurité : chiffrement transparent, possibilité de masquer des données et de brouiller le contenu de certains éléments sensibles des résultats de requêtes SQL autorisées, etc. TrustedDB [BaS11,

BaS14] propose même d'équiper les serveurs de dispositifs matériels sécurisés et impliqués dans l'exécution de manière à garantir un très haut niveau de sécurité. Ces approchent démontrent la nécessité de prendre en compte le respect de la vie privée, de réduire les risques de fuites de données et d'attaques côté serveur. Mais elles ne permettent pas de résoudre les deux problèmes intrinsèques à toute approche serveur : (1) une fuite de données ou une attaque menée avec succès compromet de très grands volumes de données et (2) le modèle étant basé sur la délégation, il peut conduire à des passe-droits et à des usages secondaires indésirables pour les usagers, incontrôlables par ces derniers.

La seconde approche consiste à offrir aux individus des serveurs personnels (réels ou virtuels) pour gérer leurs données de façon décentralisée. Cette approche est prometteuse car elle répond bien aux deux limites intrinsèques de l'approche serveur. Le projet FreedomBox[6] est l'un des pionniers à proposer une plateforme logicielle adaptée à l'architecture d'un plug computer (par exemple un Raspberry Pi) permettant aux individus de communiquer de façon anonyme et indépendante du réseau Internet classique. Les approches basées sur des serveurs personnels se démocratisent actuellement et de nombreux projets et startups proposent des solutions à destination du grand public, comme openPDS[7] [MSW+14], CozyCloud[8], Younity[9], Lima[10], OwnCloud[11], Tonido[12], Seafile[13], SparkleShare[14], etc. D'autres exemples sont discutés dans [NTB+12] et une critique principale est formulée : la difficulté de garantir à l'usager une protection contre les accès dérobés (matériel ou logiciels) potentiels et de lui garantir l'usage qui sera fait de ses données une fois celles-ci transmises hors de son serveur personnel. Mais à notre connaissance aucune de ces approches, représentatives de ce que l'on pourrait baptiser le « Web Personnel », n'intègre de composant sécurisé matériellement. Notre vision se distingue donc nettement et pourrait être une préfiguration d'un « Web Personnel Sécurisé » où l'usager pourrait avoir de fortes garanties sur ses propres données, et, n'ayant pas lui-même tous les droits sur son propre serveur, pourrait offrir des garanties à ceux qui interagissent avec lui.

---

[6] FreedomBox: http://freedomboxfoundation.org
[7] OpenPDS@MIT: http://openpds.media.mit.edu/
[8] CozyCloud: https://www.cozycloud.cc
[9] Younity: http://getyounity.com/
[10] Lima : https://meetlima.com/
[11] OwnCloud : https://owncloud.org/
[12] Tonido : http://www.tonido.com/
[13] SeaFile: http://seafile.com/en/home/
[14] SparkleShare: http://sparkleshare.org/

## 2. Approche

Notre approche est basée sur l'émergence de nouveaux dispositifs matériels sécurisés qui fleurissent actuellement sous différentes formes allant, selon le contexte applicatif, de cartes SIM multimédia nouvelle génération aux clés USB ou cartes SD sécurisées, badges d'authentification ou cartes communicantes (voir figure 1). Ces dispositifs sont présentés sous des noms différents tels que « Smart USB Token » [Eur08], « Mobile Security Card[15] » pour Giesecke & Devrient [GiD14], « Personal Portable Security Device[16] » pour Gemalto et Lexar, ou « Secure Portable Token » [AAB+10a]. Ils sont fondés sur une architecture commune combinant une puce sécurisée matériellement (ou microcontrôleur sécurisé) avec une mémoire de stockage persistante de grande capacité de type Flash NAND.

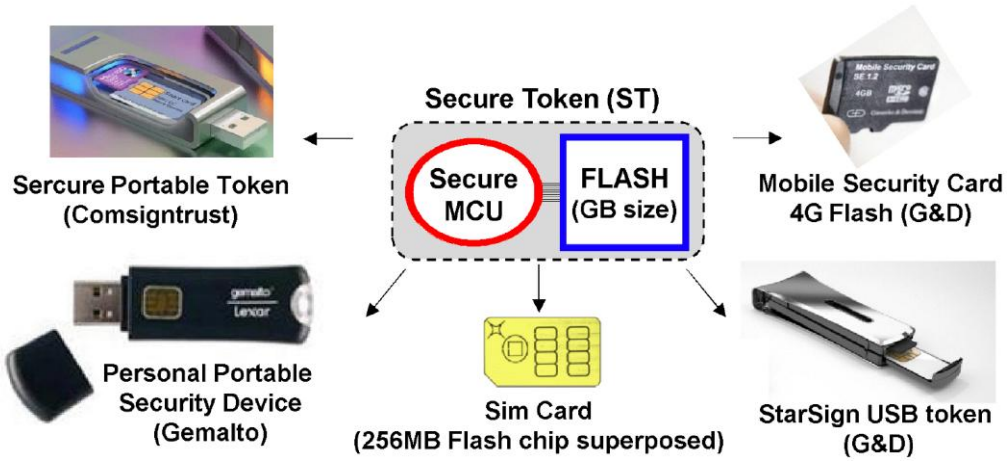


***Figure 1.*** *Exemples de dispositifs personnels et sécurisés existants*

Notre objectif est donc d'embarquer des composants logiciels permettant de collecter, stocker et partager les données personnelles d'un individu, avec des garanties tangibles de non contournement. Le dispositif offre un très haut niveau de sécurité: (1) l'attaquant a l'obligation d'être (physiquement) en contact avec le dispositif pour l'attaquer, (2) le dispositif hérite de la sécurité matérielle de la puce sécurisée qui le protège contre les attaques par canaux auxiliaires,

[15] Les produits « Mobile Security Card » de Giesecke & Devrient combinent une puce sécurisée et une mémoire de stockage de masse de type Flash NAND dans une carte microSD. Voir: http://www.gi-de.com/en/products_and_solutions/products/strong_authentication/Mobile-Security-Card-31488.jsp

[16] Voir à titre d'exemples les produits « Smart Guardian » http://cardps.com/product/gemalto-smart-guardian et « Smart Enterprise Guardian » http://cardps.com/product/gemalto-smart-enterprise-guardian

(3) le code embarqué est ouvert (open source) et peut être prouvé formellement ou certifié par la communauté ce qui le protège contre les attaques logicielles, (4) la simplicité du serveur lui permet d'être auto administré ce qui prévient la possibilité d'une attaque de l'administrateur, (5) le ratio coût/bénéfice d'une attaque comparé à un serveur classique est augmenté par les trois premiers points et par le fait qu'une attaque réussie ne compromet que les données d'un seul individu, et (6) le porteur lui-même n'a pas directement accès aux données embarquées ce qui garantit que les données obtenues provenant d'autres usagers ne seront pas compromises.

Au-delà d'un simple répertoire sécurisé de documents personnels, nous souhaitons permettre le développement de nouvelles applications orientées données et donner la possibilité à des applications existantes d'interroger le dispositif, ce qui nécessite d'organiser les documents de manière structurée, consistante et interrogeable. Nous souhaitons aussi permettre à l'usager de contrôler les règles de partage des données et lui offrir des garanties tangibles de non contournement de ces règles. Ces objectifs combinés nous conduisent à définir un véritable serveur de données, personnel et sécurisé. Les avantages d'un tel serveur sont les suivants : (1) offrir les fonctionnalités principales d'un moteur de base de données (structuration des données, contrôle d'accès, facilités d'interrogation et transactions) et être interopérable avec des sources de données existantes et avec les autres usagers, (2) permettre à l'utilisateur de contrôler le partage de ses propres données (quelles données, avec qui, pour combien de temps, à quelles fins) et garantir les principes de respect de la vie privée (consentement, collecte et rétention minimum, audit) pour ses propres données et celles appartenant à d'autres, et (3) garantir à l'usager un très haut niveau de sécurité et lui offrir un accès déconnecté aux données qu'il ne pourrait obtenir avec un serveur classique.

## 3. Contributions

L'architecture initiale que nous avons proposée se base sur une hypothèse de monde fermé, et très organisé. Elle s'adapte à certains scénarios d'usage, et sert de base à l'application DMSP présentée chapitre 4. Nous présentons ensuite une version plus ouverte de l'architecture, adaptée à la gestion de l'ensemble des données produites autour d'un individu ou d'un domicile (documents personnels, mais aussi traces de consommation électrique, données issues de capteurs domotiques, traces GPS, etc.). Enfin, nous présentons une troisième version de cette

architecture adaptée aux Pays les Moins Avancés et caractérisée par une absence d'infrastructure (faible couverture réseau, pas de serveurs centraux, pas d'autorité de certification, etc.). Ces architectures soulèvent des problèmes de recherche pour la communauté base de données dont certains font l'objet des chapitres suivants (chapitre 3 et 4).

## 3.1. Architecture « Serveur Personnel de Données » (annexe A, [AAB+10a])

L'architecture « Serveur Personnel de Données » (SPD) définit une infrastructure permettant de mettre en œuvre la vision illustrée figure 2. Les données de l'utilisateur, Bob, sont produites par différentes sources et transmises à son SPD qui peut ensuite répondre aux requêtes d'applications privées (servant les intérêts de Bob), partagées (accédant aux données de Bob depuis d'autres SPD), globales (interrogeant l'ensemble des SPD de façon anonyme) ou externes (accédant aux données de Bob sans SPD).

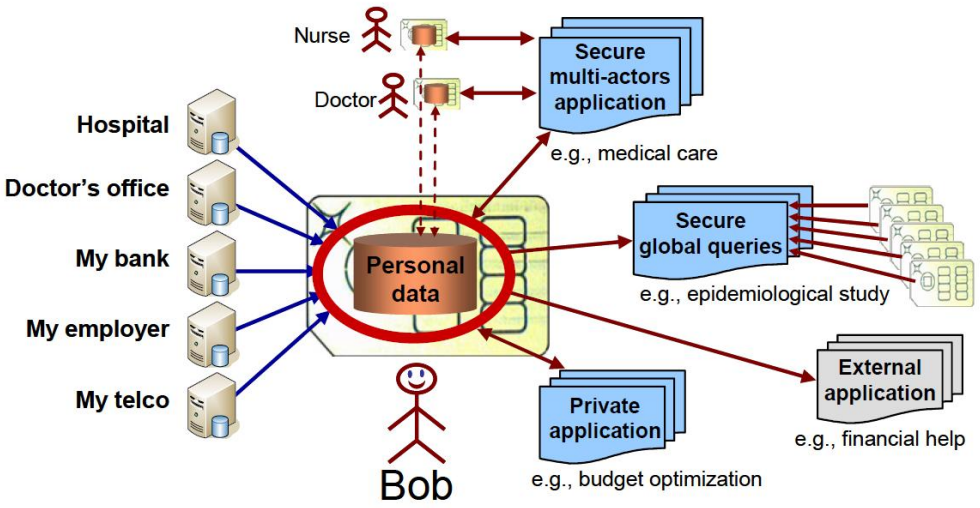


***Figure 2.*** *L'approche "Serveur Personnel de Données".*

Le SPD seul ne peut offrir toutes les fonctionnalités base de données désirées. Nous introduisons donc dans l'architecture un serveur de support responsable d'assurer la durabilité des données et de stocker les messages envoyés à destination des SPD. Ce serveur est honnête mais curieux (il effectue correctement la tâche demandée mais cherche à obtenir de l'information confidentielle), ce qui est l'hypothèse habituelle pour un service de stockage. Les données transmises au serveur de support sont donc chiffrées.

Pour que les SPD puissent interagir avec les applications, les documents stockés doivent être représentés de manière structurée et interrogeable. Dans cette proposition nous supposons que des schémas de base de données sont définis par des Fournisseurs de Schémas (des agences gouvernementales comme le ministère de la santé ou des consortiums privés comme un groupement de banques) pour chaque domaine d'application. Des Fournisseurs de Contenu fournissent des documents, en XML, suivant un format standard (comme HL7 pour des documents de santé) ou défini par un Fournisseur de Schéma. Nous considérons que chaque document (ex. une prescription médicale) est enrichi de toutes les références nécessaires (ex. les informations relatives au médecin qui a établi la prescription). Le document peut ainsi être posté vers un SPD destinataire via le serveur de support, puis téléchargé et transformé par le SPD destinataire en un ensemble de tuples de la base grâce à des règles de transformation fournies par le Fournisseur de Schéma. Ces règles sont déclaratives et vérifiables. La figure 3 illustre la transformation d'une prescription médicale transmise par un hôpital, enrichie des références aux médecins et aux médicaments. Nous supposons que la base embarquée est relationnelle mais ce choix n'a pas d'impact sur l'architecture globale.

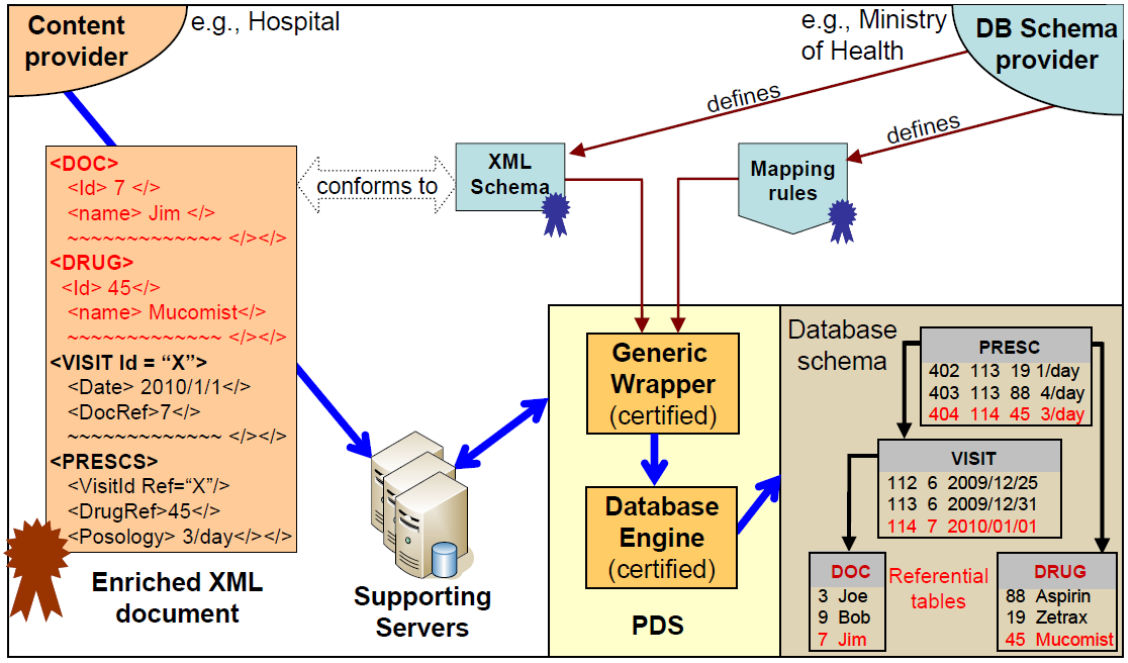


***Figure 3.*** *Insertion d'un document dans la base de données du SPD*

Les applications sont développées par des Fournisseurs d'Applications, sur les schémas publiés par les Fournisseurs de Schémas. A chaque application correspond un ensemble de *règles de collectes* spécifiant le sous-ensemble des documents requis pour son bon fonctionnement. Ces règles sont exprimées au niveau des documents pour être comprises par les usagers et sont transposées au niveau de la base de données pour être évaluées comme des règles de contrôle d'accès.

L'usager exerce un contrôle sur l'usage qui est fait de ses données en acceptant ou refusant les applications (contrairement à une application serveur classique, il peut changer d'application sans perdre ses données), il consent à ce qu'un tiers (un médecin) puisse utiliser son dossier en lui délivrant (physiquement) son SPD, qui peut identifier le médecin comme tel et limite ses droits grâce à une politique d'accès prédéfinie par le Fournisseur de Schéma (qui fixe une politique d'accès au schéma conforme à la législation pour les différentes catégories de professionnels) ou le Fournisseur d'Application. Le porteur du SPD peut aussi définir ses propres règles de masquage sur les documents de la base (et ainsi cacher des documents à certains acteurs). De plus, pour les données personnelles exportées vers un autre SPD, le « donneur » fixe des règles de divulgation minimum (durée de rétention, droits de dissémination), garde la possibilité de supprimer une donnée transmise à tout moment, et définit des règles d'audit à appliquer par le SPD destinataire pour vérifier l'usage qui est fait de ses données. Les données sont publiées auprès du serveur de support et les règles spécifiées par le donneur seront garanties par le ou les SPD destinataires.

Le serveur de support offre une zone de stockage (de données chiffrées) et un service d'horodatage (les SPD ne sont pas équipés d'horloge). Les communications sont asynchrones entre les SPD (ils sont le plus souvent déconnectés) et un service de durabilité (les SPD s'envoient des messages à eux-mêmes) permet la restauration des données d'un SPD perdu à partir d'une passe-phrase connue du porteur.

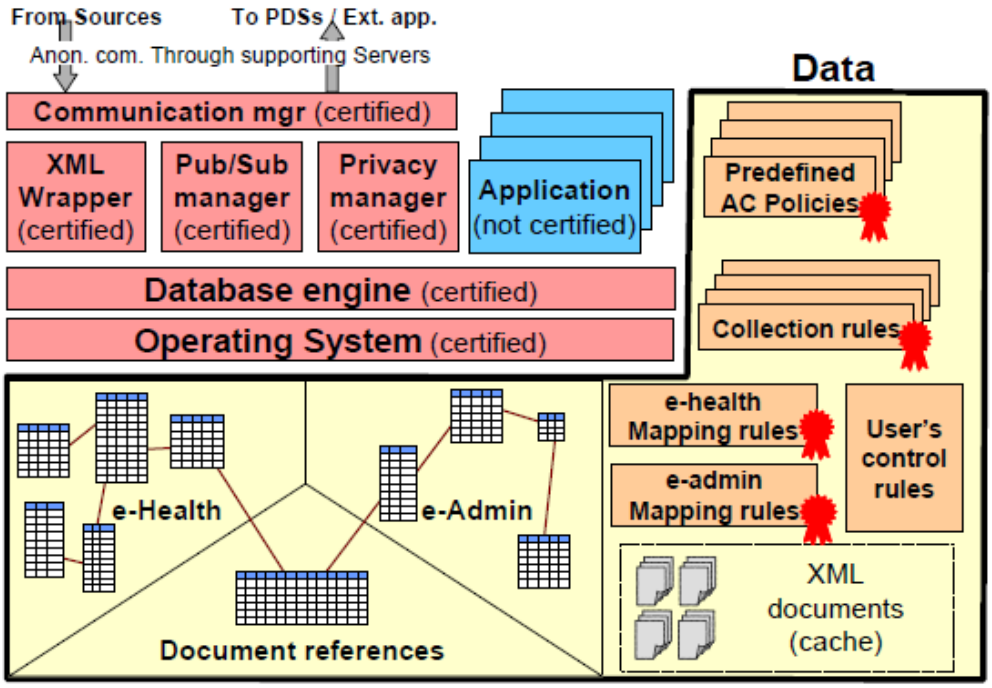


***Figure 4.*** *Logiciel générique du PDS, applications et bases de données*

La sécurité de l'architecture repose sur la sécurité matérielle du SPD, la certification du code embarqué, la ratification de règles déclaratives (règles de transformation, règles de collectes et règles de masquage) et le chiffrement de toute donnée externalisée vers le serveur de support. De plus l'anonymat des SPD se connectant au serveur de support doit être assuré au risque de révéler de l'information sensible (le volume de données transmis à un médecin peut révéler une pathologie). Les SPD intègrent un protocole rendant des communications anonymes. La certification ne concerne que certaines parties du code embarqué indiquées sur la figure 4.

### 3.2. Architecture à base de « cellules de confiance » (annexe B, [ABB+13])

Les limites de l'approche « Serveur de Données Personnel » sont liées au partage nécessairement très asynchrone car les SPD sont la plupart du temps déconnectés et aux limites imposées sur les applications qui sont embarquées dans le SDP et doivent s'adapter à de très faibles ressources. L'architecture « Cellules de Confiance » présentée dans cette section repousse ces limites et permet d'envisager des usages plus généraux.

Cette architecture se base sur les avancées récentes en matière de matériel sécurisé. AMD incorpore des processeurs de type ARM Trust Zone[17] dans ses chips visant le marché des smart phones, tablettes, boîtiers décodeurs et ordinateurs portables. Les processeurs TrustZone disposent d'un 33ème bit (matériel) sur le bus, servant à séparer matériellement les instructions provenant d'une zone « riche » du système (ouverte et dans laquelle s'exécutent les applications) d'une zone « sécurisé » (exécutant des modules de code sécurisés). TrustZone donne la possibilité de sécuriser les périphériques (ex. une partie de la RAM, les ressources d'entrées sorties comme le clavier, l'écran ou les dispositifs de stockage externe comme la carte micro SD) de manière à les rendre accessibles depuis la zone sécurisée en les isolant de la zone riche. Des entreprises comme NVidia, Sierraware ou Genode Labs permettent de rendre Trustzone utilisable depuis Linux or FreeRTOS, et Xilinx ou Trustonic proposent des plateformes matérielles de développement d'applications TrustZone [GoB14].

Cette évolution nous permet d'envisager une architecture dans laquelle de nombreux dispositifs personnels seraient constamment connectés et dotés de sécurité matérielle (figure 5). Toute donnée personnelle produite par l'espace personnel d'un utilisateur (son domicile, sa voiture, sa tablette ou son smartphone) pourrait être acheminée vers la cellule de confiance principale (fixe), par exemple intégrée ou connectée à la boîte internet du domicile. De même, des sources de données présentes dans la maison (compteur électrique intelligent, appareils domotiques) pourraient nourrir cette cellule de confiance principale.

Le SPD n'est pas pour autant absent de l'architecture. Il offre une sécurité matérielle contre les attaque physiques (et notamment les attaques du propriétaire de la cellule de confiance) que n'offre pas TrustZone. Nous voyons donc le SPD comme un composant bas niveau dans cette architecture, intégré dans la cellule Fixe et utilisé pour stocker les clés de chiffrement donnant accès aux données et évaluer les droits d'accès, alors que les ressources moins sécurisées (TrustZone) servent à exécuter des applications utilisant les données et à implanter des contrôles sur l'usage fait des données par ces applications.

[17] http://www.arm.com/products/processors/technologies/trustzone.php

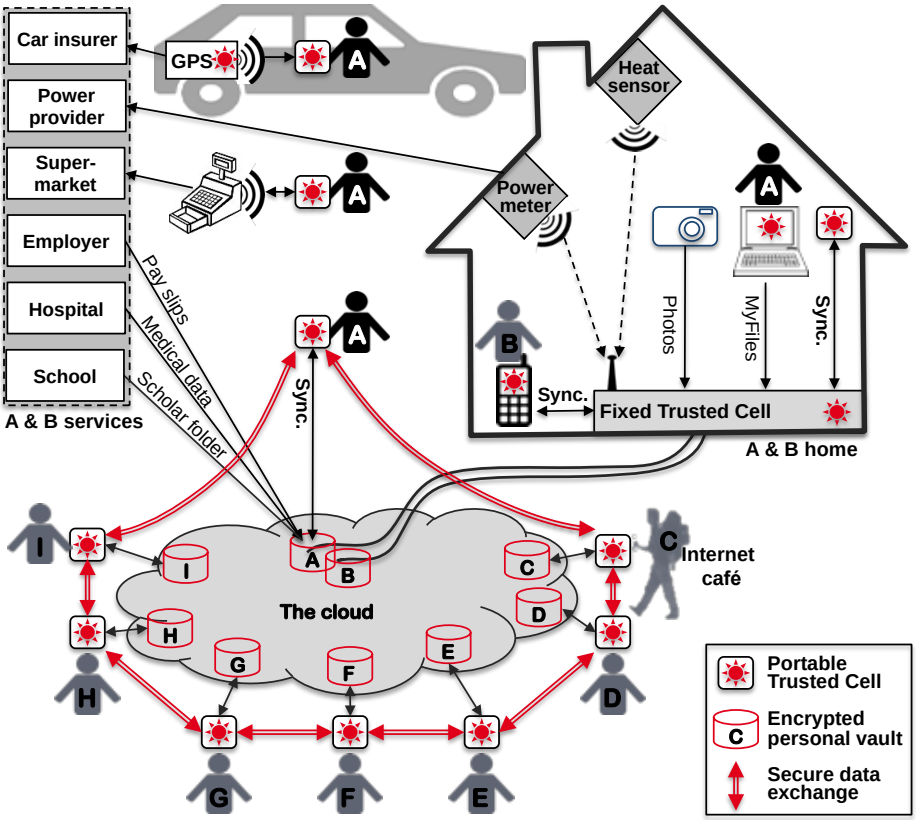


***Figure 5.*** *Alice **(A)** et Bob **(B)** sont équipés de cellules de confiance (« trusted cells ») fixes et mobiles collectant des données depuis différentes sources et les synchronisant (chiffrées) avec un espace digital personnel sur le Cloud. Tous les utilisateurs équipés de cellules de confiance peuvent partager de façon sécurisée leurs données chiffrées via le Cloud.*

### 3.3. Architecture « Folk-IS » (annexe C, [ABD+14a])

L'approche Folk-IS s'inspire de l'architecture initiale « Serveur Personnel de Données », appliquée dans le cadre des Pays les Moins Avancés (PMA). L'objectif est de fournir aux habitants un dossier numérique personnel (médical, scolaire, etc.) et la possibilité de communiquer entre eux (messages vocaux, email, etc.) et de permettre aux organisations (administrations, acteurs médicaux, ONGs, etc.) d'établir un outil d'échange des données avec les habitants (recensement de population, détection d'épidémie, programmes culturels, suivi médical et sanitaire, etc).

Le déploiement de services orientés données dans les pays les moins avancés se heurte au manque de moyens de communication (couverture 3G partielle et à coût prohibitif) et plus généralement au manque d'infrastructure technique, économique, juridique, politique et organisationnelle. Après discussions avec plusieurs ONG et acteurs locaux, nous avons identifié les besoins suivants en matière de solutions TIC: (1) le manque de sécurité institutionnelle (peu de lois protégeant les individus et peu de recours en cas d'atteinte) et de moyens d'identification (habitants sans carte d'identité) imposent à la solution TIC de garantir par elle-même la sécurité et les principes de respect de la vie privée ; (2) la solution doit présenter des bénéfices personnels immédiats car les acteurs économiques et politiques locaux n'ont pas la capacité d'imposer une solution sur le terrain ; (3) la solution doit être auto-suffisante, c'est-à-dire ne pas reposer sur une amélioration hypothétique des infrastructures ; (4) le coût par usager doit être très faible et le déploiement de la solution doit pouvoir se faire de manière incrémentale (sans investissement lourd au départ), tout en générant une source de revenus pour de nouveaux emplois locaux.

Nous proposons une solution centrée sur les habitants, basée sur l'introduction d'un composant individuel, portable, à faible coût et sécurisé, appelé *Folk-node*, capable de gérer des données personnelles et de transmettre des messages. Il s'agirait d'un composant à bas coût intégrant au minimum une puce sécurisée, une carte mémoire Flash, et un lecteur d'empreinte digitale (voir la figure 6). Les acteurs pourront accéder aux e-services (dossiers médicaux, scolaires, services personnels de messagerie) en connectant leur *Folk-node* (sans écran ni clavier) à un terminal (avec écran et clavier). Les terminaux seront détenus par des travailleurs itinérants (des personnes rémunérées pour se rendre dans les villages et partager leur terminal avec les habitants) et des acteurs locaux (enseignants, médecins, etc.). Pour permettre les échanges de données sans couverture internet, les terminaux et *Folk-nodes* seront équipés de services réseaux de routage géographique. Les messages chiffrés seront acheminés de manière transparente lors des interactions *Folk-nodes*/terminaux, en utilisant les déplacements des habitants pour parvenir de proche en proche jusqu'au destinataire qui seul pourra accéder au contenu du message. Les délais de transmission pourront être importants (plusieurs jours) mais de nombreuses applications restent compatibles avec ce type d'asynchronisme. Sans infrastructure, les déplacements des habitants et des travailleurs itinérants peuvent assurer seuls

la mise en place globale des e-services et le système peut profiter de tout élément d'infrastructure réseau existant pour diminuer la latence des échanges.

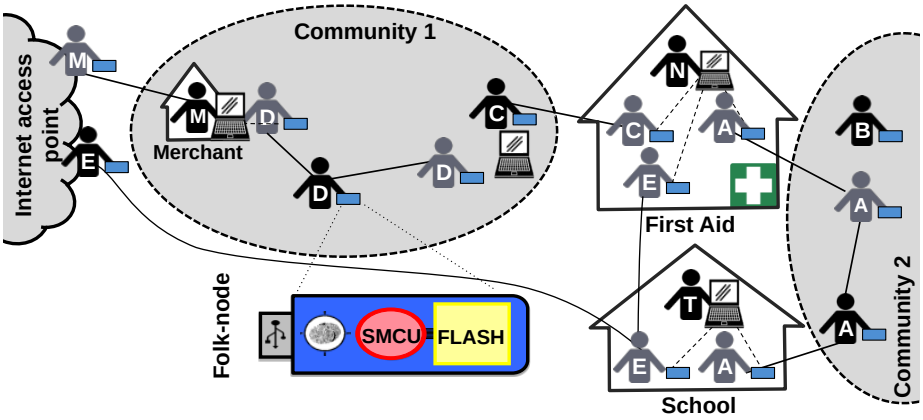


***Figure 6.*** *Deux communautés rurales, leurs habitants (icones noires) et leur déplacement (icones grises), une école et une infirmerie accessibles aux deux communautés et un point d'accès Internet. Les interactions Folk-nodes/terminaux sont représentées en pointillés. Un message transmis de A à B peut suivre différents chemins (par exemple : A → T → E → Internet) selon les déplacements des acteurs et leurs interactions avec des terminaux.*

## 4. Conclusion et résultats

Ces différentes architectures montrent que des alternatives au modèle du Web actuel peuvent être envisagées. De nombreux défis scientifiques pour la communauté base de données sont sous-jacents à ces architectures. Ils sont présentés dans les annexes A, B et C du document. Certains fondent mes travaux de recherche actuels et futurs (chapitre 5). Deux de ces défis seront abordés dans les chapitres suivants (chapitre 3 et 4) et sont transverses aux trois architectures :

***Moteur de gestion de données embarqué dans des puces sécurisées.*** Il s'agit d'embarquer dans des puces sécurisées des fonctionnalités de gestion de données assurant notamment le stockage des données, leur interrogation des données et l'évaluation de règles d'accès et de politiques de confidentialités. Le défi scientifique associé est décliné selon l'architecture cible dans les annexes (voir annexe A, section 5; annexe B, section 4, paragraphe « Secure private

store »; annexe C, section 3.1). L'étude de ce problème est présentée dans le chapitre 2 et dans l'annexe D.

***Contrôle et sécurisation de la dissémination des données de l'utilisateur.*** Ce verrou scientifique recoupe plusieurs aspects qui se posent de façon différente selon l'architecture cible (voir annexe A, section 4.3; annexe B, section 4, paragraphes « Secure sharing », « Controlled collection of sensed data » et « Secure usage and accountability » et annexe C, section 3.3). Le problème est abordé au chapitre 3 sous l'angle de la collecte minimum d'informations auprès d'un usager qui interagit avec un service externe.

De nombreuses autres perspectives de recherche sont décrites dans les annexes A, B et C. Je souhaite investiguer plus particulièrement l'une d'entre elles, liée la gestion de données sans infrastructure dans le cadre de l'architecture Folk-IS. Cette perspective est plus détaillée dans la section 2 du chapitre « Conclusion et Perspectives ».

Les résultats les plus marquants, ainsi que certaines activités de dissémination, relatifs à ces travaux sur les architectures, sont récapitulés ci-dessous.

***Articles scientifiques.*** L'architecture « SPD » initiale a fait l'objet d'un article VLDB'10 [AAB+10a]. La vision étendue aux cellules de confiance a été présentée à CIDR'13 [ABB+13], et le contexte des Pays les Moins Avancés est considéré dans un papier vision VLDB'14 [ABD+14a] et dans un papier SIGMOD Record [ABD+14b]. L'approche « Serveur Personnel Sécurisé » a donné lieu à deux tutoriaux à MDM'13 [ANP13] et à EDBT'14 [ANP14].

***Thèses.*** J'ai co-encadré les thèses de Yanli Guo [Guo11] et Lionel Le Folgoc [Fol12] qui ont toutes deux contribué au design de l'architecture « SPD ». Avec Yanli, nous avons étudié des techniques cryptographiques adaptées aux structures de la base de données embarquée et proposé des protocoles cryptographiques permettant d'assurer les communications asynchrones et anonymes entre les SPD via le serveur de support. Avec Lionel nous avons étudié des techniques de journalisation des mises à jour pour les données du SPD, efficaces en mémoire Flash et permettant d'assurer l'atomicité des transactions. Ces résultats font l'objet d'une partie de leur thèse respective et ont contribué au papier [AAB+10a] sur l'architecture.

***Plateformes matérielles et logicielles.*** Nous avons conçu une plateforme logicielle représentative de l'architecture à base de « cellules de confiance » dans le cadre du projet ANR KISS. L'architecture intègre un agent logiciel implantant une API permettant d'interfacer des applications externes avec le SPD. L'agent tourne à la fois sur systèmes Windows et Linux (PC classique et Raspberry Pi). J'ai participé à la conception de cette API, en ai supervisé une partie des développements, et ai participé à l'élaboration des tutoriaux de prise en main de la plateforme. Nous travaillons aussi en lien avec CozyCloud à la réalisation d'une plateforme de « Web personnel sécurisée » en cherchant à combiner la solution offerte par CozyCloud à la nôtre. La thèse de Paul Tran Van, dont j'ai co-encadré le stage de Master, va démarrer sur un contrat Cifre et aura pour objectif d'étudier la façon d'interfacer un composant personnel et sécurisé au « data system » de CozyCloud pour offrir de fortes garanties sur le partage et les usages des données personnelles gérées dans une instance Cozy.

***Dissémination.*** Notre objectif est de disséminer notre plateforme sous forme de logiciel libre et de matériel libre au travers de l'enseignement. La diffusons a démarré en 2014 auprès des étudiants de l'ENSIIE, dans le cadre du cours « Architectures Privacy-by-Design » que nous avons monté pour cela. L'ENSIIE a fait l'acquisition du matériel nécessaire pour mettre en place l'architecture (composants personnels sécurisés, environnement de développement et sondes matérielles de debug). Nous espérons ainsi sensibiliser les étudiants à la problématique du respect de la vie privée et leur permettre de contribuer à développer des applications innovantes et de nouveaux usages sur cette base. En 2015 nous allons disséminer cette plateforme auprès des étudiants de l'Université de Versailles St-Quentin par le biais du FabLab[18] de l'UVSQ et auprès de élèves ingénieurs de l'INSA Bourges. Nous envisageons aussi une diffusion similaire auprès des élèves ingénieurs de l'INSA Lyon. Concernant la dissémination dans la communauté scientifique, un article ERCIM News décrit une déclinaison de notre architecture telle qu'étudiée dans le cadre du projet ANR KISS [APP+12] et un autre décline cette architecture pour la ville intelligente dans le cadre du projet CityLab@Inria [ABB+14] présentée aussi à Futur en Seine 2014 [Anc14b]. Ce type d'architecture suscite un intérêt pour les acteurs du domaine publicitaire qui se voient confisquer les données

[18] http://www.fondaterra.com/projet/fablab/

personnelles par les grands acteurs du Web et qui m'ont invité à leur en faire une présentation au séminaire BIG-DATA'14 [Anc14].

# Chapitre 2

# Serveur Personnel de Données

*Ce chapitre se concentre sur le serveur personnel de données formant le cœur des architectures décentralisées présentées au chapitre précédent et vues comme une préfiguration du Web personnel sécurisé. Notre solution consiste à embarquer le code et les données dans un composant sécurisé, combinant une puce ayant un niveau de sécurité matérielle élevé (comme dans une carte bancaire) avec une mémoire Flash de grande capacité (Go). De nouveaux dispositifs personnels combinant ces deux composantes fleurissent actuellement, sous différentes formes, en fonction de leur usage applicatif. Notre objectif est ici de concevoir un moteur de gestion de données relationnel embarqué dans ce type de dispositifs. Ce chapitre résume les contraintes techniques de ces dispositifs et leur influence sur la conception du SGBD embarqué puis présente nos contributions et résultats les plus significatifs. Les contributions techniques de ce chapitre reposent sur la publication DAPD'14 [ABP+14] présentée en annexe D.*

## 1. Motivation

L'idée d'utiliser un serveur personnel, propriété de l'individu, pour gérer le patrimoine numérique de celui-ci, sous son contrôle effectif, est actuellement soutenue par de nombreux projets et startups. La plupart des initiatives allant dans ce sens se base sur des plateformes personnelles classiques (ordinateur personnel, tablette, smartphone, plug computers, etc.) pour jouer le rôle de serveur personnel. Dans le projet SMIS nous considérons des plateformes sécurisées matériellement afin de garantir à l'usager que les règles de partage et de dissémination de ses données personnelles ne pourront être contournées. Les plateformes compatibles avec cette approche combinent un microcontrôleur sécurisé avec la grande capacité de stockage des mémoires de type NAND Flash (voir chapitre 1, section 2).

Nous cherchons dans un premier temps à embarquer un SGBD relationnel dans ce type de dispositif, permettant de stocker, indexer et interroger les données en SQL, en supportant au moins les opérations de base : sélections, projections, jointures sur clé et calculs d'agrégats. Le moteur embarqué peut ainsi produire différentes vues des données personnelles avec de très fortes garanties de sécurité héritées de la sécurité physique offerte par la puce. Les données sont stockées dans une mémoire NAND Flash externe interfacée par un bus avec la puce. Un

stockage à distance est possible sur le Cloud, et dans ce cas le composant personnel gère des métadonnées (liens, attributs décrivant les données, mots clés, tags, clés de chiffrement, etc.) décrivant les données externes chiffrées (les clés de chiffrement restant confinées dans le composant personnel) et le partage de documents peut être établi en partageant ces métadonnées sous le contrôle du propriétaire des données. Le volume de données/métadonnées embarquées dans le composant personnel peut être important dès lors qu'il s'agit de gérer l'ensemble de l'histoire digitale d'un individu.

Concevoir un tel SGBD embarqué pose des difficultés techniques liées aux fortes contraintes de la puce sécurisée et de la mémoire Flash NAND. D'une part, le microcontrôleur de la puce dispose de très peu de RAM (au plus quelques dizaines de Ko) et cette quantité augmente très faiblement depuis des années[19]. D'autre part le module de Flash NAND a de très mauvaises performances d'accès lorsqu'on le soumet à de petites écritures aléatoires. De plus, ce module n'est pas dans l'enceinte sécurisée de la puce ce qui impose de chiffrer les données écrites dans cette mémoire.

Lorsqu'il s'agit de gérer de grands volumes de données, ces contraintes sont difficiles à résoudre car elles mènent à des techniques antagonistes: évaluer des requêtes base de données avec très peu de RAM conduit à indexer massivement les données pour obtenir des performances acceptables, or la maintenance de ces index lors des insertions et mises à jour induit de très nombreuses écritures aléatoires de petite taille.

## 2. Etat de l'art et formulation du problème

Les produits SGBD embarqués existants, tels SQLite ou BerkeleyDB, ainsi que les versions légères des SGBD du commerce, comme IDM DB2 Everyplace ou Oracle Database Mobile Server, visent des plateformes personnelles (smartphone ou set-top-box) relativement puissantes. Ces solutions sont clairement inadaptées aux contraintes du dispositif que nous considérons. Certains autres SGBD de l'état de l'art considèrent spécifiquement les microcontrôleurs sécurisés [PBV+01, BSS+03], mais ne considèrent pas de mémoire Flash

[19] Les ressources de la puce cohabitent sur le même dé de silicium, dont la taille doit être réduite pour apporter la sécurité matérielle désirée. Or, la RAM a une faible densité, ce qui conduit les industriels à favoriser les autres composants (notamment la quantité de mémoire stable).

externe. Ils sont adaptés à de faibles volumes de données stockées dans des mémoires internes (technologie Flash NOR ou EEPROM) avec des caractéristiques très différentes de la Flash NAND (notamment en termes de granularité des accès). Les techniques pensées dans ce contexte ne peuvent être transposées à notre cas.

Certaines techniques classiques en bases de données permettent d'obtenir de bonnes performances lorsque les requêtes SQL nécessitent d'évaluer des jointures ou des calculs d'agrégats sur de grands volumes de données par rapport à la quantité de RAM disponible pour effectuer le calcul. Mais les performances des algorithmes classiques de jointure (par boucle imbriquée, par tri-fusion, par hachage de Grace ou hybride) se détériorent rapidement[20] lorsque la taille du plus petit argument de la jointure dépasse la taille de la RAM [HCL+97]. Des algorithmes plus récents comme le « Jive Join » et le « Slam Join » utilisent des indices de jointure [LiR99] mais nécessitent tout de même une taille de RAM de l'ordre de la racine carrée de la taille de la plus petite des tables impliquée dans la jointure. Dans notre contexte, le ratio entre la taille de la RAM et celle des tables est si petit que la seule solution est de considérer un modèle très fortement indexé, où toutes les jointures (au moins sur clé) sont pré calculées, comme dans le cas d'un entrepôt de données. Par exemple, pour évaluer une jointure en étoile impliquant une très grande table des Faits, les entrepôts de données indexent habituellement la table des Faits sur toutes ses clés étrangères, ce qui revient à pré calculer la jointure avec toutes les tables Dimensions et sur tous les attributs de ces tables participant à la requête [Sun99, Wei02]. Cependant une indexation aussi massive nécessite un très grand nombre d'écritures aléatoires de petite taille lors des insertions des données afin de maintenir les index à jour, ce qui dans notre contexte induirait un coût inacceptable lié aux écritures en Flash NAND.

Les problèmes de la gestion de données en mémoire Flash NAND et de sa couche de traduction ont fait l'objet de nombreux travaux de la communauté bases de données. Ce type de mémoire supporte très mal les écritures aléatoires de petite taille. La mémoire est en effet divisée en blocs contenant chacun des pages (ex. 64), elles-mêmes découpées en secteurs. La granularité d'écriture est la page (ou le secteur) et les écritures doivent être réalisées séquentiellement dans un même bloc. Une page ne peut être réécrite sans que le bloc contenant

[20] Avec 10Ko de RAM, joindre des tables de 100Ko par boucle imbriquée conduit à lire la seconde table 10 fois (par bloc), puis la troisième 100 fois, etc.

cette page n'ait été effacé au préalable. Le nombre d'effacements possibles de chaque bloc est borné (ex. $10^4$ effacements). Habituellement ces contraintes sont masquées par un module de traduction optimisant l'accès à la mémoire Flash qui intègre une couche de traduction d'adresses permettant de réaliser les mises à jour hors place, un ramasse miette pour réclamer les blocs dont les données sont devenues obsolètes et un mécanisme de nivellement de l'usure des blocs permettant un effacement équitable. De nombreux travaux (voir [KoV11]) proposent des améliorations du module de traduction. Cependant, avec très peu de RAM, le module de traduction ne peut pas masquer les contraintes de la Flash de manière efficace. Dans notre contexte le microcontrôleur sécurisé peut disposer d'un accès direct au composant Flash (s'il est soudé au dispositif) ou d'un accès via une couche de traduction (si le composant Flash est intégré dans une carte SD ou microSD). Dans le premier cas les techniques de traduction efficaces consomment de la RAM, ce qui proscrit leur utilisation dans notre contexte. Dans le second cas les écritures aléatoires par page (ou secteur) sont beaucoup plus coûteuses que les écritures séquentielles (les temps d'écritures aléatoires par page ou secteur de 100 à 1000 fois plus coûteux[21] qu'en séquentiel[22]).

De nombreuses études récentes proposent des solutions au problème du stockage et de l'indexation en mémoire Flash NAND. Les index classiques, comme l'arbre B+, se comportent très mal lorsqu'ils sont implantés en mémoire Flash au-dessus d'une couche de traduction [WCK03]. Les solutions actuelles adaptent en général l'arbre B+ en journalisant les mises à jour pour les intégrer par batch dans l'arbre et minimiser le nombre d'écritures aléatoires en Flash. Le journal des mises à jour doit être indexé en RAM pour assurer des performances satisfaisantes. Les différentes propositions diffèrent par la façon de gérer le journal et l'index en RAM, et par l'impact de cette gestion sur la fréquence de l'intégration du contenu du journal dans l'arbre B+. Pour obtenir un gain important sur le temps d'écriture en Flash, le journal doit être intégré très peu fréquemment dans l'arbre B+, mais cela conduit à consommer plus de RAM. Inversement, minimiser la consommation RAM conduit à intégrer le journal à l'arbre B+

[21] Des tests sur 20 cartes SD récentes montrent les écritures aléatoires sont en moyenne 1300 fois plus couteuses que les écritures séquentielles [SmR]. Nous obtenons des ratios de l'ordre de 100 à 1000 sur les cartes SD et microSD dont nous disposons dans l'équipe.

[22] Cela n'est pas le cas des disques Flash SSDs, qui disposent d'une quantité de RAM interne relativement large (ex. 16 MB), et peuvent offrir un coût d'écriture aléatoire proche de celui d'une écriture séquentielle.

plus fréquemment, et donc à plus d'écritures aléatoires. Avec la très faible quantité de RAM disponible dans un microcontrôleur, la fréquence d'intégration des mises à jour est de fait élevée, donnant un gain minime sur les écritures aléatoires.

D'autre travaux proposent des index basés sur des structures séquentielles [Arg03, MOP+00, OCG+96], inspirés des systèmes de gestion de fichiers organisés sous forme de journaux [RoO92] ou de fichiers différentiels [SeL76] où une zone séquentielle sert à stocker les nouveaux enregistrements qui seront intégrés à terme dans le système de gestion données principal. Ces systèmes utilisent aussi de grands tampons en RAM, incompatibles avec nos contraintes. Plus récemment le système Hyder a été proposé [BRD11] pour gérer sous forme séquentielle une base de données clé valeur en mémoire Flash. Ce système utilise un (seul) arbre binaire pour retrouver l'enregistrement correspondant à une clé donnée. A chaque mise à jour de l'arbre, au lieu de modifier l'arbre en place, le chemin complet depuis la racine jusqu'à l'élément inséré ou modifié est réécrit. Mais cette technique n'est pas adaptée à un système massivement indexé, car les arbres binaires sont adaptés aux index sur clés uniques mais pas aux index secondaires. D'autres propositions de systèmes clé valeur en Flash, comme SkimpyStash [DeS11], LogBase [VWA+12] ou SILT [LFA11], organisent les paires clé valeur dans des journaux pour éviter les écritures aléatoires, et maintiennent en mémoire (RAM) des index d'une taille proportionnelle à celle de la base de données (au minimum 1 octet par enregistrement), ce qui est incompatible avec les contraintes d'un microcontrôleur.

Pour conclure, concevoir un SGBD, avec des performances acceptables, sur des Go de données, avec une toute petite RAM et une grande mémoire Flash NAND, reste un problème ouvert. Les solutions actuelles sont contradictoires, nécessitant d'indexer massivement la base de données (petite RAM) avec pour conséquence d'engendrer de très nombreuses mises à jours aléatoires pour maintenir les index (coûteuses en Flash) sauf à consommer plus de RAM pour amortir ces coûts. Notre objectif est de proposer des techniques permettant de rompre ce cercle vicieux.

## 3. Approche

Notre solution doit répondre aux objectifs contradictoires suivants : (1) indexer massivement la base de données, (2) produire exclusivement des écritures séquentielles en Flash et (3) consommer une toute petite quantité de RAM indépendante de la taille de la base de données.

Nous proposons pour cela d'organiser toute la base de données (les données, index, tampons, journaux transactionnels, etc.) dans des structures de données purement séquentielles que nous appelons des *Containers Séquentiels* (CS). Un CS satisfait trois conditions : (1) son contenu est écrit séquentiellement dans les blocs de Flash qui lui sont alloués (une fois écrites, les pages ne sont jamais modifiées ni déplacées); (2) de nouveaux blocs peuvent être alloués pour étendre le CS; (3) un CS est libéré entièrement lorsqu'il devient obsolète (pas de libération partielle de l'espace du CS).

L'intérêt d'adopter une stratégie purement séquentielle est d'éviter les écritures aléatoires par définition. Cependant les traitements appliqués sur les structures séquentielles ne passeront pas à l'échelle. Pour permettre la gestion de grands volumes de données, la base de données séquentielle initiale devra être réorganisée de manière itérative dans de nouvelles structures plus performantes. Cette nouvelle organisation devant être elle aussi produite dans des CS pour satisfaire l'objectif de départ.

## 4. Contributions [ABP+14]

Nos contributions se situent (1) au niveau de l'évaluation de requêtes (sélection, projection, jointure, calculs d'agrégats) avec une petite quantité de RAM bornée, (2) au niveau de l'organisation de la base de données sous forme de CS notamment pour gérer les tampons, l'atomicité de la base de données, les mises à jour et les index de sélection et de jointure, et (3) au niveau de la protection cryptographique des données contenues dans la Flash (qui n'est pas protégée matériellement comme l'est le microcontrôleur). Les sections suivantes résument les deux premières contributions, la troisième est décrite dans [ABP+14] et les détails sont donnés dans [Guo11].

### 4.1. Stratégie d'évaluation de requête avec une petite RAM

Avec une très faible quantité de RAM par rapport au volume de données à traiter, nous devons considérer des index généralisés de sélection et de jointure [ABB+09, PBV+01, Sun99, Wei02] qui capturent toutes les relations directes et transitives entre les tuples. Sur cette base un index généralisé peut être défini avec deux types d'index. Le premier type d'index que nous appelons *TJoin* (pour Jointure Transitive) pré calcule la jointure naturelle d'une table

référençant d'autres tables (directement ou transitivement). Ainsi, un index TJoin construit sur la table $T_i$ vers la tables $T_j$, noté $I_{Ti \rightarrow Tj}$, associe à chaque tuple $t$ de la table $T_i$ le tuple $t'$ de la table $T_j$ référencé par $t$ (pour lequel il existe dans la base un chemin de jointures sur clés de $t$ vers $t'$). Le second type d'index que nous appelons *TSelect* (pour Sélection Transitive) pré calcule une sélection sur les valeurs d'un attribut d'une table référencée (directement ou transitivement) par une autre table. Un index TSelect construit sur la colonne $T_j.A$ d'une table $T_j$ vers une autre table $T_i$ qui référence $T_j$, noté $I_{Tj.A \rightarrow Ti}$, associe à chaque valeur $v$ de la colonne $T_j.A$ tous les tuples de $T_i$ qui référencent un tuple de $T_j$ pour lequel la valeur de $A$ est $v$. Les résultats doivent être triés pour permettre les unions et intersections des listes sans consommer de RAM (par simple fusion de listes triées).

Par exemple, sur le schéma de la base de données considéré sur la figure 7, les références entre les tables sont indiquées par les flèches en pointillés : la table $T_0$ référence directement $T_1$ et $T_3$ et référence transitivement $T_2$, $T_4$ and $T_5$. Dans cet exemple, 8 index TJoin sont créés, représentés par les flèches pleines de la figure 7.a. Les flèches pleines de la figure 7.b présentent les index TSelect, construits sur les attributs $a$, $b$, $c$, $d$, $e$, et $f$ des tables $T_0$ à $T_5$, considérant qu'un (seul) attribut de chaque table $T_i$ est indexé. Pour l'attribut $c$ de la table $T_2$, nous créons donc 3 index TSelect: $I_{T2.c \rightarrow T2}$, $I_{T2.c \rightarrow T1}$, $I_{T2.c \rightarrow T0}$.

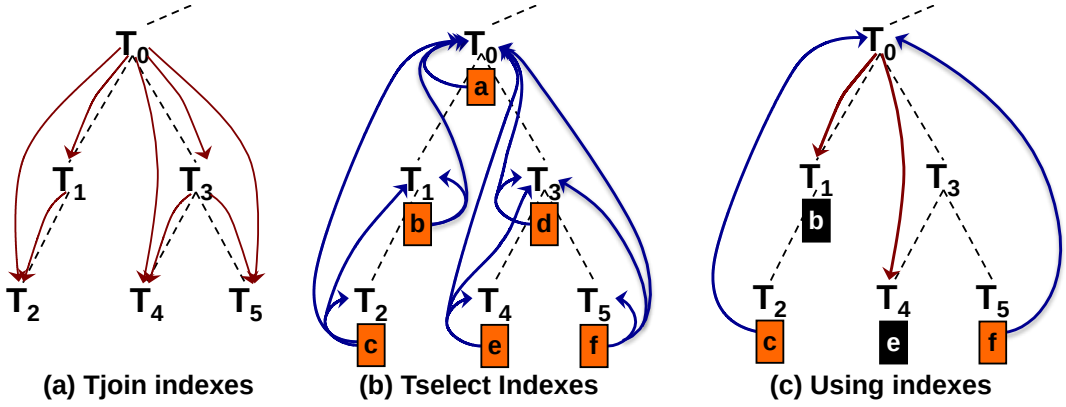


***Figure 7.*** *Exemple d'un schéma d'indexation massif et son utilisation.*

Intuitivement, une requête impliquant des sélections, projections et jointures (SPJ) est évaluée en : (1) traversant les index TSelect construits sur les attributs impliqués dans une sélection (l'index $I_{Tj.A \rightarrow Ti}$ est impliqué si la requête contient un prédicat de la forme $T_j.A$ et si $T_i$

est la table commune de tous les index TSelect impliqués dans la requête), (2) fusionnant les ensembles triés d'identifiants de tuples de $T_i$ en pipeline (en appliquant une intersection et/ou une union); et (3) traversant les index TJoin nécessaires pour projeter les tuples résultats. La figure 3.c montre le procédé appliqué pour évaluer une requête joignant toutes les tables avec le prédicat de sélection ($T_2.c = v_1$ *et* $T_5.f = v_2$) et projetant les attributs $T_1.b$ and $T_4.e$. Cela conduit à : (1) accéder à $I_{T2.c \rightarrow T0}(v_1) \rightarrow S_1$ et $I_{T5.f \rightarrow T0}(v_2) \rightarrow S_2$, (2) faire l'intersection des ensembles triés $S_1$ et $S_2$ en pipeline et (3) utiliser les index $I_{T0 \rightarrow T1}$ et $I_{T0 \rightarrow T4}$ pour retrouver les tuples résultats et projeter les attributs $T_1.b$ et $T_4.e$.

Les requêtes avec une clause de groupement sont plus difficiles à évaluer car la consommation RAM est liée au nombre de groupes résultats. Dans ce cas, le résultat de la partie SPJ de la requête est stocké dans un CS temporaire puis toute la RAM est utilisée pour calculer les agrégats en plusieurs itérations sur ce CS, produisant une fraction du résultat à chaque itération. Nous proposons diverses stratégies d'optimisation pour ce traitement dans [ABB+09].

## 4.2. Organisation séquentielle de la base de données

***Tampons et atomicité.*** Les tuples des tables de la base peuvent facilement être organisés séquentiellement sous forme d'un ensemble de CS nommé ↓DATA, en adoptant une organisation des tables soit en ligne, soit en colonne. L'insertion de nouveaux enregistrements produit des données à ajouter (séquentiellement) aux CS de ↓DATA. Des tampons doivent être utilisés pour stocker les mises à jour à grain fin (potentiellement plus petites que la taille d'une page Flash) jusqu'à obtenir une page pleine à ajouter à un CS donné. Ils sont eux aussi organisés sous forme d'un ensemble de CS nommé ↓BUF et sont utilisés non seulement comme tampon pour tous les autres CS de la base, mais servent aussi à garantir l'atomicité des mises à jour. La bonne gestion des tampons et de l'atomicité de la base de données, afin de minimiser le nombre d'écritures global dans les LC, a fait l'objet d'une partie du travail de thèse de Lionel Le Folgoc [Fol12].

***Mises à jour.*** Les mises à jour (respectivement, les suppressions) d'enregistrements ne peuvent être reportées directement dans les CS (par définition), mais sont journalisées dans un ensemble de CS dédiés nommé ↓UPD (resp. ↓DEL). Pour gérer les mises à jour, les anciennes

et nouvelles valeurs des attributs mis à jour sont journalisées dans ↓MAJ. Lors de l'exécution des requêtes le moteur vérifie dans ↓UPD si des mises à jour peuvent impacter leur résultat. Si une mise à jour correspond à un prédicat de requête, la requête est compensée en éliminant les faux positifs (enregistrements qualifiés pour le prédicat de la requête sur leur ancienne valeur mais pas sur la nouvelle) et en intégrant les faux négatifs (enregistrements qualifiés sur leur nouvelle valeur mais pas sur l'ancienne). Ensuite, ↓UPD et ↓DEL sont vérifiés à nouveau lors de l'étape de projection des résultats afin de projeter les nouvelles valeurs et de retirer du résultat les enregistrements supprimés. Le surcoût est minimisé en indexant les structures ↓UPD et ↓DEL en Flash et en maintenant en RAM des résumés partiels de ces structures sous forme de filtres de Bloom. La stratégie proposée est simple mais certains détails d'implémentation plus arides sont exposés dans [Fol12].

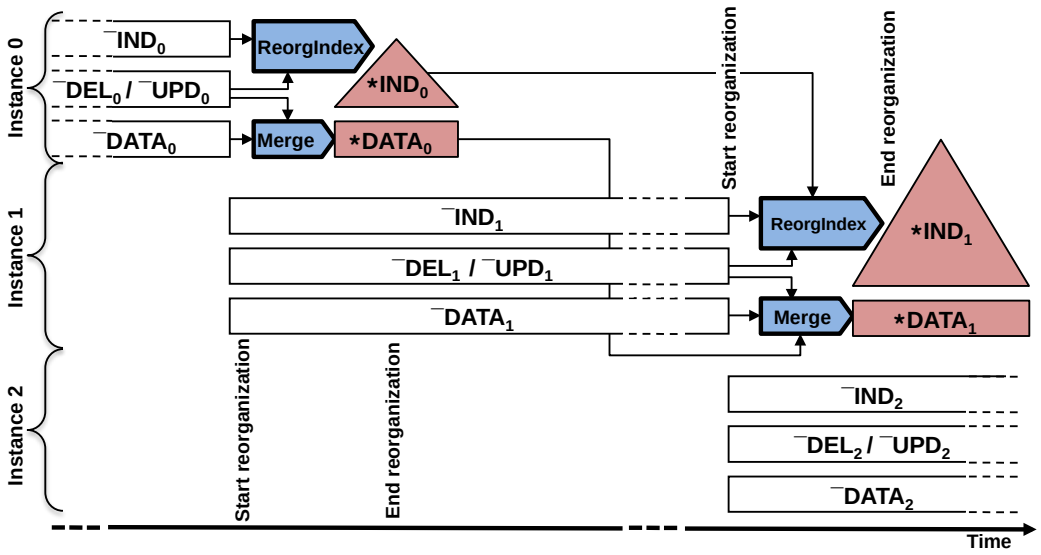


***Figure 8.*** *Le processus de réorganisation.*

***Index séquentiels.*** Bien que quelques structures d'indexation existantes soient naturellement compatibles avec la notion de CS (ex. l'index bitmap), la plupart des index sont incompatibles avec cette notion (ex. les structures arborescentes et celles basées sur du hachage) car l'insertion de nouvelles données génèrerait des mises à jours dans des nœuds ou paquets de hachage existants. Nous avons donc proposé de nouvelles formes d'index séquentiels compatibles avec la notion de CS et applicables aux index TSelect et TJoin nécessaires à notre stratégie

d'évaluation de requête avec une petite RAM. Ces structures sont décrites en détail dans [ABP+14]. Elles sont dans la suite notées ↓IND.

***Passage à l'échelle.*** Pour gérer efficacement de grands volumes de données nous transformons les structures séquentielles de la base de données initiale, notées ↓DB, en une base de données dite **optimale*, générée elle aussi séquentiellement, notée *DB. La base *DB est dite **optimale* dans le sens où elle permet d'obtenir des performances d'interrogation aussi bonnes que si elle avait été construite avec les méthodes de l'état de l'art en ignorant les contraintes de mise à jour de la Flash. Par exemple, la réorganisation de ↓DB en *DB peut donner un ensemble de tables dans lesquelles toutes les modifications et suppressions sont intégrées, et un ensemble d'index organisés à base de structures arborescentes ou basées sur du hachage. Notons que le processus de réorganisation que nous proposons est indépendant du schéma de stockage et d'indexation sous-jacent. Avant toute réorganisation nous avons $\downarrow DB_0$ = ($\downarrow BUF_0$, $\downarrow DATA_0$, $\downarrow IND_0$, $\downarrow UPD_0$, $\downarrow DEL_0$), l'indice 0 représentant un compteur incrémenté à la fin de chaque réorganisation. Lorsque $\downarrow DB_0$ atteint une limite (en termes de taille ou de performance en interrogation), il faut construire $*DB_0$. La réorganisation déclenche trois actions décrites figure 8 : (1) le contenu de $\downarrow BUF_0$ est intégré dans les CS cibles de $\downarrow DATA_0$, $\downarrow IND_0$, $\downarrow UPD_0$ et $\downarrow DEL_0$, (2) $\downarrow DB_1$ est alloué et toutes les nouvelles insertions, modifications et suppressions sont redirigées vers $\downarrow DB_1$, et (3) $\downarrow DB_0$ (mise en lecture seule) est réorganisée en $*DB_0$, composée de $*DATA_0$ et $*IND_0$. $*DATA_0$ est construit en intégrant dans $\downarrow DATA_0$ toutes les modifications et suppressions journalisées dans $\downarrow UPD_0$ et $\downarrow DEL_0$ (opération *Merge*). $*IND_0$ est la réorganisation **optimale* de $\downarrow UPD_0$, intégrant les éléments de $\downarrow UPD_0$ et $\downarrow DEL_0$ (opération *ReorgIndex*, cette opération de réorganisation des index est décrite dans [ABP+14]). La base de données est alors composée de $\downarrow DB_1$ (utilisé pour stocker les nouvelles données) et de $*DB_0$ composé de $*DATA_0$ et $*IND_0$. A la fin du processus de réorganisation (lorsque $*DATA_0$ and $*IND_0$ sont construits complètement) tous les CS de $\downarrow DB_0$ sont libérés. La taille de $\downarrow DB_1$ augmente ensuite jusqu'à atteindre une limite qui déclenchera la prochaine phase de réorganisation. La réorganisation est ainsi un processus itératif très différent des approches journalisant les mises à jour pour ensuite les intégrer par batch qui produisent toujours des écritures aléatoires.

## 5. Conclusion et résultats

MiloDB est le premier serveur embarqué dans un microcontrôleur sécurisé à même de considérer de grands volumes de données stockées dans une mémoire Flash NAND et supportant l'ensemble de l'algèbre relationnelle. Le haut degré de sécurité est obtenu grâce aux 3 propriétés suivantes : (1) la protection matérielle du microcontrôleur; (2) l'embarquement du code et son évaluation permettant de n'externaliser que des résultats autorisés sur les données ; et (3) le stockage chiffré des données dans la mémoire Flash NAND. Cette étude a généré des résultats scientifiques, logiciels, et sert de base à des actions de dissémination, les plus marquants étant récapitulés ci-dessous.

***Articles scientifiques.*** Certains éléments du modèle d'exécution de requêtes embarqué ont été d'abord introduits dans un contexte plus simple, où seules certaines colonnes sensibles étaient embarquées et étaient non modifiables, les autres colonnes restant stockées sur un serveur public. Cette étude a fait l'objet des publications SIGMOD'07 [ABB+07], DAPD'09 [ABB+09] et d'une démonstration à VLDB'07 [SAB+07]. L'étude présentée dans ce chapitre a consisté à embarquer toute la base dans le composant personnel, en supportant les mises à jour, ce qui a fait l'objet d'une démonstration à SIGMOD'10 [ABG+10] et d'un article journal DAPD'14 [ABP+14].

***Thèses.*** Les thèses de Yanli Guo [Guo11], Lionel Le Folgoc [Fol12] et Saliha Lallali (en cours) considèrent la conception de techniques de gestion de données pour un serveur embarqué. Avec Yanli nous avons proposé des techniques de protection cryptographique des données et des index adaptées au stockage en Flash et aux motifs d'accès et structures du SGBD embarqué. Avec Lionel nous nous sommes concentrés sur l'organisation des tampons et sur l'atomicité transactionnelle. Leurs travaux de thèse ont contribué aux publications [ABG+10, ABP+14]. Avec Saliha, nous étudions actuellement la généralisation des techniques proposées pour le relationnel à d'autres modèles (clé-valeur, recherche de documents, etc.).

***Logiciels.*** Le logiciel embarqué, nommé *PlugDB-engine*, est développé sur la base du design présenté ici. Il a fait l'objet de trois dépôts successifs à l'Agence de Protection des Programmes [ABP+08, ABP+09a, ABP+11] et un nouveau dépôt est en cours pour fin 2014. J'en pilote les

développements qui impliquent depuis 2007 des doctorants et ingénieurs de l'équipe SMIS. Ce logiciel est utilisé dans de nombreuses activités de l'équipe décrites dans ce document : l'application DMSP (voir chapitre 4), les projets CG78/DMSP, PlugDB et KISS (voir Introduction, section 4), le module SIPD1 et 2 à l'ENSIIE (chapitre 1, section 4), etc.

***Dissémination et transfert.*** Un contrat de transfert du logiciel PlugDB-engine est actuellement en cours de signature avec la startup CozyCloud qui propose une solution de Cloud personnel. Le transfert vise à interfacer le « data system » de CozyCloud avec PlugDB-engine de manière à maintenir des fichiers chiffrés dans l'instance Cozy de l'usager, et les clés de chiffrement correspondantes dans PlugDB-engine, associées à des métadonnées décrivant les fichiers et sur lesquels définir des règles d'accès. En couplant la solution Cozy avec PlugDB-engine, nous souhaitons offrir aux usagers une première version d'un « Web Personnel Sécurisé » décrite en introduction du manuscrit.

Nos perspectives de recherche consistent à embarquer toujours plus de fonctionnalités de gestion de données dans le serveur personnel pour pouvoir réaliser en local le plus de calculs possibles et n'externaliser que les résultats de ces calculs, et pour être en mesure d'évaluer des règles d'accès plus riches dans l'enceinte sécurisée. Ces perspectives sont détaillées dans la section 2 du chapitre « Conclusion et Perspectives ».

# Chapitre 3

# Exposition Minimum

*Ce chapitre montre comment le serveur personnel d'un usager peut permettre une minimisation de l'exposition des données d'un individu lors d'une interaction avec un service externe. Nous nous plaçons dans le cadre de la collecte de données via des formulaires telle qu'elle est pratiquée par les organisations (aide sociale, banques, assurances, administration, etc.) souhaitant ajuster leur offre à la situation spécifique d'un demandeur. Ce chapitre introduit le contexte et le mode opératoire actuel, puis présente nos contributions techniques dans le cas de décisions modélisables par des classifieurs multi-labels (ensembles d'arbres de décision) et conclut par un résumé de nos résultats principaux. Les contributions techniques du chapitre reposent sur la publication [ANV12a] présentée en annexe E.*

## 1. Contexte

La directive *EU95/46/CE* [Dir95] fait référence en Europe concernant la protection des données à caractère personnel et prône l'application de principes de collecte (et de rétention) minimum des données personnelles. Leur application permet notamment de réduire les effets d'une fuite de données postérieure à la collecte (causée par une attaque, une négligence ou un usage détourné de la finalité de la collecte). Cette directive stipule que les données collectées doivent être « ***non excessives*** *au regard des finalités pour lesquelles elles sont collectées* ». Suite au scandale PRISM révélant les accès de la NSA aux données des usagers collectées par Google, Facebook, Microsoft, Apple, etc., la Commission Européenne procède actuellement à une refonte de cette directive. Le projet actuel [Res14], adopté en première lecture le 12 mars 2014, renforce ces principes en stipulant que les données personnelles doivent être « *limitées au* ***minimum*** *nécessaire au regard des finalités pour lesquelles elles sont traitées* ».

L'application du principe de collecte minimum participe aussi à réduire les coûts pour les organisations. Le traitement des données et leur archivage nécessitent parfois des ressources importantes et l'intervention d'employés. Les coûts engendrés dépendent grandement de la quantité d'information à traiter. Par exemple, les demandes d'aide sociale sollicitées auprès des Conseils Généraux se font au travers de formulaires très complets comme le formulaire GEVA (« Guide d'EVAluation » des besoins de compensation de la personne dépendantes), découpés

en plusieurs volets (médical, social, ressources financières, entourage, etc.) et comportant des dizaines de pages et des centaines de champs. Les demandes sont nombreuses (60.000 par an pour le seul Conseil Général des Yvelines) et impliquent beaucoup d'employés (160 personnes au CG78) chargés de vérifier l'information renseignée, de calibrer l'aide à apporter puis d'archiver les demandes (rejetées et acceptées) pour pouvoir justifier la décision en cas de contestation, en attester la nature non discriminante et permettre des audits réguliers.

Dans la pratique, les formulaires à renseigner sont construits par les organisations en faisant l'union de tous les attributs pouvant avoir un impact sur la décision finale d'offre de service. Pourtant, pour un individu donné, seul un sous ensemble de l'information est pertinent. Par exemple, une personne dépendante pourrait bénéficier d'une aide financière pour employer une aide journalière à domicile dans les cas suivants : (i) sa pension de retraite annuelle est inférieure à 30.000€ et elle a plus de 80 ans, (ii) sa pension de retraite est inférieure à 10.000€ quel que soit son âge, ou (iii) son score Groupe Iso-Ressources (GIR) indiquant son niveau de dépendance (sur une échelle de 1 à 6) est supérieur à 2. Pour une personne $p_1$ = [*pension* = 25.000, *age* = 81, *GIR* = 1] l'ensemble minimum à produire serait [*pension*, *age*]. Pour une personne $p_2$ = [*pension* = 40.000, *age* = 60, *GIR* = 3] il suffirait de produire [*GIR*]. Cet exemple simpliste montre que le contenu minimum du formulaire ne peut être produit a priori mais dépend des valeurs des attributs demandés pour la personne concernée.

Notre objectif est de proposer des techniques permettant la minimisation des données à renseigner dans le formulaire et respectant les hypothèses suivantes : (1) l'offre de service obtenue à partir du formulaire minimisé doit être identique à celle qui aurait été obtenue à partir du formulaire complet et les valeur des attributs du demandeur justifiant l'offre de service doivent pouvoir être connues (et archivées) par l'organisation ; (2) le processus de décision ne doit pas être exposé au demandeur car il peut révéler le modèle d'affaire (« business model ») de l'organisation (prêts bancaires, contrats d'assurance) ou inciter à la fraude (falsification de certaines valeurs d'attributs pour obtenir des bénéfices supplémentaires) ; notre solution doit aussi être (3) adaptée à des formulaires de grande taille, rencontrés couramment dans le cas de l'attribution d'aides sociales ou lors des interactions avec les banques, les compagnies

d’assurance, ou administration fiscale ; enfin (4) le principe proposé doit pouvoir convenir pour un large spectre de processus de décision.

## 2. Etat de l’art

La transposition de principes légaux dans les systèmes informatiques a été la base de nombreux travaux au cours de la dernière décennie. Des exemples emblématiques incluent la plateforme P3P [CLM+02], les langages de définition de politiques de vie privée comme EPAL [AHK+03] ou encore les bases de données Hippocratiques [AKS+02]. La technologie P3P permet de mettre en évidence des problèmes d'incompatibilité de politiques de confidentialité entre un individu et un service mais ne permet pas de choisir un sous ensemble des données à exposer. Les langages de définition de politiques de confidentialité qui ont été proposés, comme EPAL, XACML [Mos05] ou WSPL [And04], n'ont pas non plus été introduits dans l’objectif de minimiser la collecte des données. En revanche l'architecture d'une base de données Hippocratique repose sur dix principes de respect de la vie privée tirés de la législation qui incluent bien la collecte limitée des données. Un SGBD Hippocratique limite la collecte d’informations en associant à chaque objectif de traitement l’ensemble des attributs requis pour atteindre cet objectif. Toutefois cette solution fait l'hypothèse que les données utiles et inutiles au traitement pour atteindre l’objectif peuvent être déterminées en amont de la collecte. Comme nous l'avons montré section 1, c'est peut-être vrai dans certains cas simples mais cela n'est en général pas le cas des processus de décision complexes.

Un domaine reposant sur des techniques proches de celles utilisées dans notre étude est le domaine de la négociation automatique de confiance, notamment dans le contexte des modèles de contrôle d'accès ouverts où les décisions d'accès résultent d’une confrontation entre les politiques de contrôle d'accès des parties et les valeurs d’attributs qui les caractérisent. Un petit nombre de travaux suit une approche de collecte minimale nommée dans ce contexte *exposition minimale* [ADF+12, CCK+05, YFA+08]. Ces travaux minimisent le nombre de valeurs d’attributs à exposer par les parties lors de la phase de négociation automatique pour permettre la prise de décision (donner ou pas l'accès à une ressource). Toutefois le problème et les solutions sont différents pour deux raisons essentielles. Tout d'abord les processus de prise de décision que nous considérons sont plus complexes que ceux sous-jacents au contrôle d'accès.

Les règles de collecte que nous considérons modélisent des classifieurs multi labels (ensemble d'arbres de décision) car de nombreuses dimensions peuvent être considérées (ex. pour une demande de prêt bancaire : taux plus faible, durée plus longue, assurance réduite, etc.) dont chacune peut impacter l'offre finale proposée à l'utilisateur. D'autre part, dans notre contexte, la prise de décision nécessite de très grandes quantités de données personnelles (les formulaires contiennent souvent des centaines de champs) tandis que dans le domaine du contrôle d'accès, seules quelques autorisations sont prises en compte (ex. jusqu'à 35 dans les évaluations de performance présentées dans [ADF+12]). Ces travaux ne peuvent donc pas être réutilisés dans notre contexte puisqu'ils ne sont pas pertinents en termes d'expressivité et de passage à l'échelle.

## 3. Approche

Le procédé actuel pour calibrer une offre de service à la situation particulière du demandeur, illustré figure 9, se déroule selon les étapes suivantes : (1) le demandeur récupère le formulaire d'application vierge correspondant à sa demande ; (2) il le renseigne conformément à sa situation personnelle et le transmet à l'organisme concerné ; (3) l'organisme vérifie la validité des informations transmises (en utilisant des preuves de provenance ou en croisant les données avec des informations dont elle dispose) et détermine l'offre de service correspondante. L'organisme (4) archive le formulaire et la décision correspondante et (5) transmet la décision à l'intéressé.

Notre approche pour réduire le formulaire tout en restant conforme à cette pratique est basée sur trois ingrédients : des règles de collectes modélisant le processus de décision de l'organisme, les données du demandeur modélisées de manière adaptée aux règles de collectes et une métrique d'exposition des données permettant d'évaluer la quantité d'informations personnelles présentes dans un formulaire donné. De plus notre solution doit aussi permettre de confronter les règles de collectes et les données du demandeur sans divulguer ni les règles au demandeur ni les données (n'apparaissant pas dans le résultat) à l'organisme. Pour cela, un algorithme de minimisation de l'exposition doit déterminer parmi les données du demandeur, le sous ensemble minimum permettant d'offrir au demandeur tous les avantages auxquels il a droit et cette minimisation doit se faire dans l'enceinte sécurisée du serveur personnel.

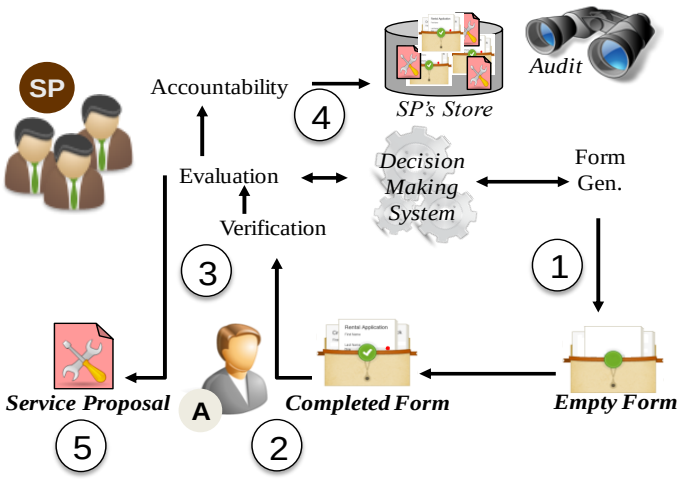


***Figure 9.*** *Architecture classique de collecte de données par formulaire.*

## 4. Contributions [ANV12a]

Nos contributions se situent au niveau : (1) de l'architecture à collecte minimum à base de serveur personnel sécurisé ; (2) de la modélisation du problème passant par une représentation des règles de collecte, des données de l'utilisateur et de la métrique d'exposition permettant d'appliquer des algorithmes calculant le contenu minimum du formulaire pour le demandeur ; et (3) de la résolution du problème et de la validation de nos solutions sur des cas d'usage réels. Les sections suivantes résument chacune de ces contributions principales.

***Architecture à collecte minimum.*** L'architecture à base de serveur personnel que nous considérons est présentée figure 10. Deux étapes additionnelles sont introduites au procédé actuel, les autres étapes restant inchangées. D'abord des règles de collecte sont construites de manière à refléter (tout ou partie) du processus de décision. Ces règles sont transmises au demandeur avec le formulaire d'application. La transmission des règles vers le serveur personnel du demandeur passe par un canal sécurisé classique (type SSH) pour éviter que le demandeur ne puisse accéder aux règles en clair. Ensuite les règles de collecte et les données du demandeur sont confrontées dans l'enceinte sécurisée du serveur personnel (à laquelle n'a accès aucune des parties) en appliquant un algorithme de réduction de l'exposition des données à renseigner dans le formulaire.

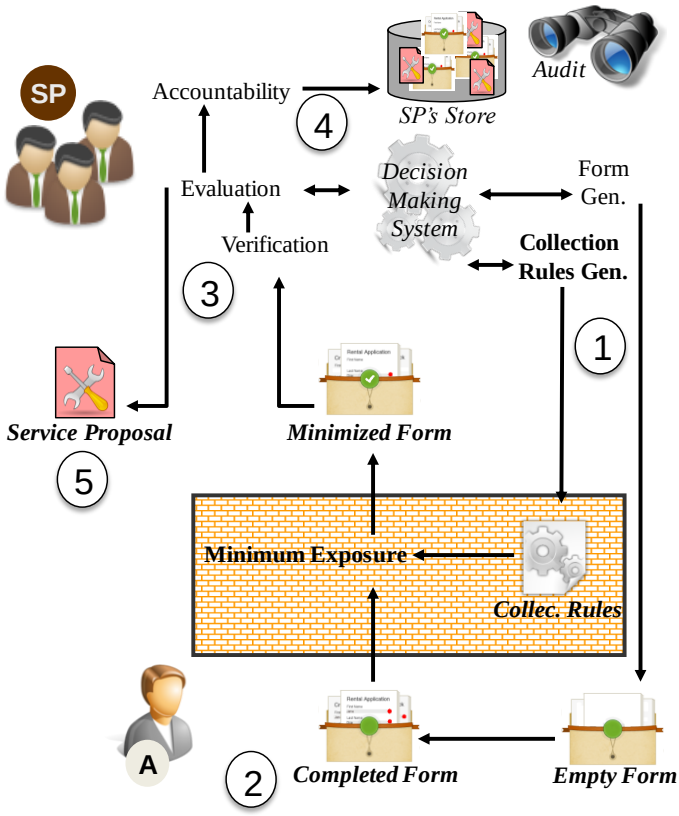


***Figure 10.*** *Architecture à collecte minimum.*

***Modélisation du problème.*** La modélisation du problème que nous proposons a juste pour but de démontrer l'applicabilité de l'approche dans certains cas d'usage réels. Nous considérons que les données personnelles d'un individu sont représentées par un ensemble de couples (attribut, valeur). Comme métrique d'exposition, nous considérons qu'une valeur est attribuée à chaque couple dénotant sa sensibilité[23] et nous calculons l'exposition d'un ensemble de couples en faisant la somme des valeurs de sensibilité des couples de l'ensemble[24]. Pour simplifier l'explication, nous considérerons dans la suite que tous les couples (attribut, valeur) ont une

[23] Une bonne métrique d'exposition de l'ensemble des couples présents dans un formulaire devrait permettre de satisfaire à la fois les exigences de l'usager (respect de sa vie privée) et l'objectif de l'organisme (réduire les coûts de traitement du formulaire), mais c'est probablement un problème de recherche en soi, dépassant l'objectif que nous nous donnons ici.

[24] Notons que certaines métriques existantes mesurant la perte d'information créées dans le cadre de travaux sur l'anonymisation de données, comme la *distorsion minimale* [XiT06] ou *ILoss* [YFA+08], peuvent elles aussi être supportées dans notre formalisme.

valeur de sensibilité de 1 ce qui revient à évaluer la sensibilité d'un formulaire comme étant le nombre d'attributs différents renseignés. Concernant les règles de collecte, nous exprimons chaque bénéfice potentiel sous la forme d'un ensemble de disjonctions de conjonctions de prédicats sur les attributs (forme normale disjonctive) permettant d'attribuer ce bénéfice au demandeur. Ce formalisme permet de représenter les décisions basées sur des règles logiques, des arbres de décision et des forêts d'arbres de décision, représentant ainsi de nombreux classifieurs binaires, multi-classes et multi-labels [TsK07]. Par exemple la règle de collecte permettant d'obtenir une aide financière donnée en exemple en introduction du chapitre, peut s'écrire :

$aide\_financière$ : ($pension < 30.000 \wedge age > 80$) $\vee$ $pension < 10.000 \vee GIR > 2$

Un ensemble de couples (attribut, valeur), lorsqu'ils sont exposés dans le formulaire, valident donc certaines règles de collecte qui déterminent les bénéfices à attribuer au demandeur. Par exemple, le couple (*GIR*, 3) valide le prédicat $GIR > 2$ et donc la règle de collecte *aide_financière*. Le problème, dit de *n-exposition*, que nous considérons, est ainsi le suivant : d'après un ensemble de règles de collecte $R$ et un ensemble de données personnelles $P=\{(a,v)\}$, $P$ est $n$-exposable pour $R$ si et seulement si il existe un sous ensemble de $m \leq n$ éléments dans $P$ qui valide toutes les règles de $R$. Minimiser le formulaire revient à trouver l'ensemble $n$-exposable ayant la valeur minimum de $n$.

***Complexité du problème.*** Le problème de *n-exposition* est NP-complet et le problème d'optimisation associé (trouver la valeur minimum de $n$) est NP-difficile. En effet le problème du *MIN-SAT pondéré*, connu pour être NP-complet, est réductible au problème de *n-exposition*. Le problème du *MIN-SAT pondéré* est formulé comme suit dans [5] : soit un entier $n$, un ensemble $\{P_{j,k}\}$ de variables booléennes, une formule logique sous forme normale conjonctive $F = \wedge_j ( \vee_k P_{j,k} )$ sur $\{P_{j,k}\}$, et une fonction pondérée positive $w:\{P_{j,k}\} \rightarrow R^+$, trouver un assignement $T$ de valeurs de vérité pour $\{P_{j,k}\}$ tel que $F$ soit satisfaite et $w(T) = \Sigma_{j,k}\, w(P_{j,k}) \times T(P_{j,k})$ soit inférieure à $n$. Pour un individu donné, en ne considérant que l'ensemble $R$ des règles de collecte satisfaites par les données de cet individu et réduites aux « conjonctions satisfaites », le problème de $n$-exposition peut être formulé de façon similaire. La satisfaction de l'ensemble $R$ des règles de collecte satisfaites pour l'individu est représenté par une formule logique $F' = \wedge_j$ (

$\vee_k$ ( $\wedge_l$ $P_{j,k,l}$ ) où les $P_{j,k,l}$ sont des variables booléennes dont la valeur est *vraie* si le couple (attribut, valeur) validant le prédicat correspondant de la règle de collecte est exposé par l'individu, et *fausse* s'il n'est pas exposé. La fonction *F*, exprimée dans le cas du *MIN-SAT pondéré*, se ramène à la fonction *F*' de la *n*-exposition, en prenant *l*=1 dans *F*' (cas d'une fonction *F*' qui représenterait uniquement des règles de collecte purement disjonctives). Les résultats de complexité obtenus pour le *MIN-SAT pondéré* s'appliquent donc aussi à la *n*-exposition. Notamment le problème du *MIN-SAT pondéré* est NP-complet et le problème d'optimisation relatif est NP-difficile [Co71, Ka72]. De plus le problème d'optimisation du *MIN-SAT pondéré à valeurs positives* n'est pas dans la classe APX [AAG+98] (on ne peut pas trouver d'algorithme d'approximation de la solution fonctionnant en temps polynomial et dont l'approximation serait bornée par une constante). Ces résultats s'appliquent également à la *n*-exposition.

***Algorithmes de résolution.*** Le problème de *n*-exposition, représenté sous forme logique, peut être résolu de manière exacte en utilisant un solveur de programmation en nombres entiers existant. Nous avons utilisé le solveur COUENNE [BLM+09], distribué en open source par le projet COIN-OR[25] qui est l'un des plus connus et efficaces parmi les solveurs gratuits adaptés à notre problème. Dans notre architecture un tel solveur ne peut toutefois pas être envisagé car, d'une part il ne trouve la solution dans un temps acceptable que pour de petites instances du problème (formulaires de quelques dizaines de champs), et d'autre part il est beaucoup trop consommateur de ressources pour pouvoir être embarqué dans l'environnement contraint du serveur personnel sécurisé[26]. Nous avons donc proposé des algorithmes de résolution heuristiques embarqués dans le serveur personnel.

Utiliser le solveur COUENNE nécessite d'exprimer le problème d'optimisation de *n*-exposition sous forme d'un programme en nombres entiers. Nous avons choisi le langage de programmation AMLP [FGK02], un langage de modélisation algébrique de problèmes d'optimisation supporté par COUENNE. Le problème de *n*-exposition exprimé sous forme de problème SAT peut être transposé en AMPL en utilisant une variable binaire par couple

[25] Voir http://www.coin-or.org/Couenne/

[26] Notons que faire tourner le solveur hors de l'enceinte sécurisée conduirait à divulguer les règles de collecte en clair hors du matériel sécurisé, ce qui serait contraire aux hypothèses que nous nous sommes fixées.

(attribut, valeur) impliqué, en exprimant une contrainte pour chaque règle de collecte et en choisissant comme fonction objectif la somme des variables binaires du problème. Un exemple simplifié est donné figure 11.

**Règles de collectes :**

$r_1$: $(p_1 \wedge p_2) \vee (p_3 \wedge p_4) \Rightarrow c_1$
$r_2$: $(p_5 \wedge p_6 \wedge p_7) \vee (p_4 \wedge p_8 \wedge p_9) \Rightarrow c_2$
$r_3$: $(p_1 \wedge p_6 \wedge p_7) \vee (p_2 \wedge p_4 \wedge p_{10}) \Rightarrow c_3$
$r_4$: $(p_2 \wedge p_5 \wedge p_6 \wedge p_7) \vee (p_1 \wedge p_4 \wedge p_8 \wedge p_9) \Rightarrow c_4$

***avec* :**

$p_1$: *pension < 30.000*, $p_2$: *age > 80*,
$p_3$: *tutelle = 1*, $p_4$: *GIR > 2*,
$p_5$: *vit_seul = 1*, $p_6$: *entourage = 0*,
$p_7$: *mobilité = 0.5*, $p_8$: *traitement = 1*,
$p_9$: *isolement > 0.5*, $p_{10}$: *plain_pied = 1*.

$c_1$=*aide_financière*, $c_2$=*assistante_journalière*, $c_3$=*adaptation_lmogement*, $c_4$=*portage_repas*.

**Formulaire à renseigner :**

*Pension, age, tutelle, GIR, vit_seul, entourage, mobilité, traitement, isolement, plain_pied.*

**Données de l'utilisateur :**

(*pension, 25.000*), (*age, 83*),
(*tutelle, 1*), (*GIR, 3*),
(*vit_seul, 1*), (*entourage, 0*),
(*mobilité, 0.5*), (*traitement, 1*),
(*isolement, 1*), (*plain_pied, 1*).

**Programme AMPL :**

```
var b1 binary; ... var b10 binary;
minimize EX:
b1+b2+b3+b4+b5+b6+b7+b8+b9+b10;
subject to
r1: b1*b2 + b3*b4 >= 1;
r2: b5*b6*b7 + b4*b8*b9 >= 1;
r3: b1*b6*b7 + b2*b4*b10 >= 1;
r4: b2*b5*b6*b7 + b1*b4*b8*b9 >= 1;
```

**Solution minimale :**

(*pension, 25.000*), (*age, 83*),
∅ ∅,
(*vit_seul, 1*), (*entourage, 0*),
(*mobilité, 0.5*), ∅,
∅, ∅.

***Figure 11.*** *Exemple de problème de n-exposition et formulation AMPL.*

Nous avons implanté plusieurs algorithmes donnant une solution approchée de façon adaptée aux contraintes du serveur personnel. L'algorithme appelé RAND* choisit plusieurs solutions de façon aléatoire et conserve la meilleure. Pour trouver une solution, cet algorithme tire au hasard pour chaque règle de collecte l'une des conjonctions de prédicats formant la règle et expose les couples (attribut, valeur) validant les prédicats impliqués dans cette conjonction. La solution validant l'ensemble des règles est l'union des couples choisis pour chaque règle. Le procédé est répété et la meilleure solution est conservée d'une exécution sur l'autre. D'autres algorithmes ont été testés, basés sur la méthode du recuit simulé ou sur des heuristiques adaptées aux scénarios réels que nous avons étudiés. L'heuristique qui donne globalement les meilleurs résultats sur les scénarios réels testés sélectionne pour chaque règle la ou les conjonctions contenant le moins de prédicats et dont la validation minimise le nombre de prédicats à valider dans les règles restant à traiter. La complexité de l'algorithme est supérieure à celle de RAND* mais pour faire une comparaison équitable nous avons calibré le nombre de

solutions considérées par RAND* de manière à donner aux deux algorithmes le même temps d'exécution.

***Résultats expérimentaux.*** L'objectif des mesures est de pouvoir valider l'approche et de quantifier les gains obtenus sur la réduction de l'exposition des données dans des scénarios réels. Le cas d'usage réel que nous avons le plus investigué concerne l'attribution d'aide sociale par les Conseils Généraux sur la base du formulaire GEVA. Divers bénéfices sont attribués au demandeur à partir de ce formulaire, et couvrent différentes dimensions: aide financière, médicale, para médicale, aide humaine (ménagère, repas, compagnie), adaptation du logement, etc. Nous avons modélisé les décisions sous forme de règles de collectes en partenariat avec le Conseil Général des Yvelines. Les décisions se basent à la fois sur la régulation en vigueur et sur des choix propres aux décisionnaires du département. Les critères de décision ne doivent pas être explicitement présentés aux usagers pour éviter d'inciter à la fraude et parce que la décision finale d'attribution de l'aide reste souveraine. Nous avons identifié, en interaction avec les services du Conseil Général des Yvelines, 63 règles de collecte, impliquant 440 prédicats. Parallèlement nous avons construit des classifieurs multi-labels sur deux jeux de données réels publics : ENRON (un ensemble d'emails devenus publics suite au scandale ENRON) et MEDICAL (rendu public par le département de radiologie d'un hôpital du Cincinnati). Ces jeux de données sont sans rapport avec le cas des formulaires d'application qui nous intéressent, mais les règles de collecte obtenues à partir de ces données suivent une topologie différente de celle que obtenue à partir du formulaire GEVA et nous permettent de tester les algorithmes sur d'autres topologies. Nous avons appliqué différentes techniques de résolution (exactes, heuristiques) et avons tiré les conclusions suivantes : (1) la réduction d'exposition est (presque toujours) conséquente avec les algorithmes heuristiques, apportant une réduction d'exposition des données renseignées dans le formulaire variant de 30% à 80% selon la topologie du problème ; (2) la résolution exacte n'est pas souvent possible (elle prend plusieurs heures dès que le nombre d'entrées dépasse 50 à 100 selon les topologies) ; (3) l'algorithme heuristique présenté ci-dessus, exécuté dans le microcontrôleur du serveur personnel, donne des résultat approchés satisfaisants proches de ceux de COUENNE (quelques points d'écart sur le pourcentage de réduction pour les instances sur lesquelles la résolution exacte est possible) et meilleurs que ceux obtenus avec l'algorithme aléatoire RAND* à temps d'exécution comparable.

## 5. Conclusion et résultats

Ces travaux nous permettent de démontrer l'intérêt du serveur personnel dans le contexte de la collecte minimum de données. Les règles de collecte et les données peuvent être confrontées dans l'enceinte sécurisée du microcontrôleur, préservant ainsi la confidentialité des règles et des données. Les algorithmes heuristiques embarqués présentent de bons résultats sur les cas réels, en termes de réduction d'exposition et de temps d'exécution (moins d'une minute pour les instances les plus grandes). De nombreux processus de prise de décisions peuvent être couverts par le formalisme que nous proposons. Les résultats scientifiques, logiciels, et les actions de dissémination les plus marquants sont récapitulés ci-dessous.

***Articles scientifiques.*** Nous avons introduit le concept d'exposition minimum comme interprétation stricte du principe légal de collecte limitée dans [ANV12a, ANV13]. Nous avons conduit une étude expérimentale utilisant des données réelles et des classifieurs multi-label pour démontrer l'applicabilité des techniques d'exposition minimum [ABN+15]. Une démonstration de techniques d'exposition minimum embarquées dans un composant à microcontrôleur et leur application au cas de l'attribution d'aides sociales a été présentée à EDBT'13 [ABN+13].

***Logiciel.*** Le prototype MinExp-Card a été développé dans le cadre du projet ANR KISS en lien avec le Conseil Général des Yvelines. Le prototype implante les techniques heuristiques d'exposition minimum sur une carte de développement STMicroelectronics STM32L152-EVAL (microcontrôleur ARM Cortex-M3, RISC 32 bit avec 16KB de RAM et 128KB de stockage persistant). Ce prototype démontre que ces techniques peuvent être portées sur des cartes à puce bon marché (quelques euros la carte) ou sur les serveurs personnels sécurisés considérés dans nos travaux.

***Dissémination.*** Les travaux sur l'exposition minimum ont été présentés dans le cadre de Workshop [ANV11, ABN+12]. Nous avons aussi écrit un article dans le magazine « Tangente » [AnB14], destiné aux élèves de lycées, dans lequel nous présentons le problème de l'exposition minimum, et proposons au lecteur un défi sur un jeu de règles de collectes inspiré de celui construit dans le cas de l'aide sociale.

Un article de conclusion, identifiant certaines attaques à l'exposition minimum et proposant des contremesures, est actuellement en préparation. En effet, la connaissance des règles de collecte, de l'objectif de l'algorithme ou de l'algorithme lui-même, peut conduire l'organisme destinataire du formulaire, ou un autre tiers, à inférer un certain nombre de données personnelles du demandeur non présentes dans le formulaire minimisé. Ces attaques peuvent être prises en compte. D'une part, les algorithmes proposés peuvent intégrer un calcul d'inférence qui complètera les formulaires minimisés avec les données inférées afin de ne pas fausser le calcul de la mesure d'exposition. D'autre part nous espérons pouvoir produire de nouveaux algorithmes de n-exposition, capables de considérer leur propre inférence à l'exécution pour produire des résultats offrant une réduction plus importante.

Dans nos travaux futurs nous chercherons à élargir le principe de l'exposition minimum au contexte du contrôle d'usage. Cette piste de recherche est décrite plus en détail dans la section 3 du chapitre « Conclusion et Perspectives ».

# Chapitre 4
# Application DMSP

*Les sections précédentes mentionnent certaines des applications qui ont motivé les travaux de recherche présentés. L'application « Dossier Médico-Social Personnel » DMSP est particulièrement emblématique pour le projet SMIS car elle repose sur l'architecture « Serveur de Données Personnel » et a été expérimentée sur le terrain. Elle donne un poids supplémentaire à nos arguments vers un « Web Personnel Sécurisé » auprès de la communauté non scientifique, industrielle et du grand public. Nous coordonnons le projet CG78/DMSP avec Philippe Pucheral et j'en pilote les développements depuis 2007. Ce chapitre présente la motivation de l'application, l'état de l'art, l'approche suivie et les principaux résultats obtenus.*

## 1. Motivation

Le vieillissement de la population impose d'améliorer le suivi sanitaire des personnes dépendantes à domicile. Dans ce contexte, des informations médicales, sociales et administratives doivent être échangées entre les acteurs intervenant dans la prise en charge (médecins, aide-ménagères, aides-soignants, assistantes sociales, auxiliaires de vie, kinésithérapeutes, etc.). Cette coordination passe naturellement par un accès à ces données au chevet du patient ou lorsque celui-ci se rend en consultation dans le cadre de son suivi médico-social et également à distance hors de la présence du patient (par exemple pour des prises de décision par le praticien interrogé au téléphone).

S'agissant de données de santé ou de données sociales, chacun des intervenants doit avoir des droits d'accès différenciés aux données. La personne suivie, avec l'aide de son médecin traitant ou de son entourage, doit pouvoir consentir (ou non) à ce que certains professionnels jouent un rôle sur son dossier. De plus le patient doit pouvoir masquer certaines données particulièrement sensibles avec l'aide de son médecin traitant. Ceci permet de faire face à des situations humainement complexes, comme par exemple un patient se sachant en fin de vie et ne voulant pas le dévoiler pour des raisons humaines et/ou financières, ou désirant ne pas révéler une pathologie à ses proches.

Les différents intervenants du circuit médico-social disposent, en règle générale, de leurs propres logiciels informatiques : logiciel de cabinet médical, de service hospitalier, logiciel infirmier, de coordinations gérontologiques, etc. Il serait donc souhaitable que les données des dossiers patients puissent se synchroniser avec ces outils pré existants afin d'éviter les doubles saisies.

## 2. Etat de l'art

Concernant la gestion des dossiers médicaux, trois approches principales se distinguent. La première consiste à interconnecter des systèmes autonomes pré existants dans une infrastructure régionale ou nationale avec un contrôle central minimal selon l'exemple danois (*Medcom*) ou nord-américain (*eHealth Exchange, NHIN*). Une deuxième approche renforce l'intégration grâce à des index (ou des résumés de données) centralisés, à l'image du projet Néerlandais (relancé en 2013 après avoir été abandonné pour la défiance qu'il inspirait aux patients) ou du projet autrichien (ELGA). La troisième approche est totalement centralisée, comme en témoignent le système VistA développé aux USA et le projet DMP national français.

Dans la pratique la coordination des soins à domicile s'effectue souvent aux travers d'un dossier papier conservé au domicile des personnes suivies. Par exemple, l'ALDS a mis au point un « Dossier Médical Commun » papier, permettant aux intervenants de reporter les faits importants du suivi des personnes dépendantes. Un intercalaire est disponible dans ce dossier pour chaque type d'intervenant lui permettant de consigner les faits marquants survenus et de les partager avec les autres intervenants. Une feuille générale « tableau de bord » permet aussi de porter toute indication significative.

Quelques solutions informatisées de coordination pour la prise en charge à domicile apparaissent. Par exemple, la société Arcan[27] (groupe Chèque Déjeuner) propose des solutions logicielles mobiles pour la coordination de soins à domicile et lance actuellement une application pour Windows phone. La solution Globule[28] présente le même type de

[27] Voir : http://www.arcan.fr/

[28] Voir : http://www.globule.net/fr/index.html

fonctionnalités. Ce genre d'application centralise l'information de coordination sur un serveur central et la rend accessible depuis des applications mobiles. Ces applications nécessitent en général un accès réseau pour avoir accès au dossier centralisé. Certaines permettent parfois de continuer à fonctionner en mode déconnecté, permettant par exemple de saisir certaines données à domicile même hors de toute couverture réseau (les données seront synchronisées sur le serveur dès que le mobile retrouvera un accès réseau). Ces solutions ont un défaut majeur : elles nécessitent que tous les professionnels gravitant autour du patient s'équipent d'un même outil logiciel. Dans la pratique, les structures qui interviennent autour d'un même patient sont très nombreuses (aide sociale diligentée par la mairie, le département, une société privée, acteurs médicaux et sociaux provenant de divers organismes de coordination gérontologique, de cliniques, installés en cabinet, etc.) et toutes ne peuvent s'équiper d'un même outil. Ce type de solution conduit donc à des dossiers partiels (beaucoup moins complets qu'un dossier papier conservé au domicile du patient) et conduit à des doubles saisies (de nombreux organismes ayant déjà leur propre outil informatique). De plus, ces solutions ne remplissent pas les critères de sécurité habituels en matière de gestion de données de santé (identification forte impossible depuis les smartphones qui ne sont pas dotés d'un lecteur de carte CPS, fonctions de respect de la vie privée moins évidentes car il s'agit avant tout d'un dossier de professionnel auquel les patients et l'entourage n'ont pas accès). Enfin ces systèmes peuvent conduire à une défiance de certains intervenants (et même des personnes suivies ou de leurs proches) qui peuvent se sentir en situation de surveillance.

## 3. Approche

Notre approche s'inspire de l'architecture « Serveur Personnel de Données » (voir chapitre 1, section 4.1). Elle consiste à équiper chaque patient d'un SPD embarquant son dossier médico-social, capable d'authentifier chacun des intervenants, de lui donner accès à la vue du dossier correspondant à sa pratique et de se synchroniser sans connexion internet avec un serveur distant permettant la sauvegarde du dossier et sa synchronisation avec différents logiciels ou chaînes de traitement externes.

Pour permettre la synchronisation entre le SPD et le serveur central sans connexion Internet, nous utilisons le matériel des intervenants (leur lecteur de carte CPS ou leur

tablette/smartphone) pour convoyer des paquets chiffrés entre le SPD et le serveur central accessible en zone connectée. Pour permettre l'interopérabilité avec des logiciels ou chaînes de traitement externes, des connecteurs adaptés peuvent être créés en partenariat avec certains éditeurs logiciels. Plus généralement le SPD peut produire un fichier, suivant un standard établi, que tout logiciel de coordination pourrait à terme reconnaître et savoir intégrer à l'image de ce qui se pratique aux Etats-Unis dans le cadre de l'initiative « Blue Button[29] ».

Le patient reste maître de ses données et est dépositaire de l'ensemble de son dossier. La complétude est acquise de fait car tous les intervenants nourrissent le même dossier, celui du patient. Ce dernier peut protéger la confidentialité de son dossier en habilitant certains professionnels à y accéder à distance (l'habilitation est automatique lorsque le professionnel se rend au chevet du patient) et il peut masquer certaines données perçues comme très sensibles avec l'aide de son médecin traitant. La puce sécurisée lui garantit que les droits d'accès associés à chaque intervenant par les autorités sanitaires et sociales ainsi que ses propres règles de masquage et habilitations, ne pourront pas être contournées.

## 4. Résultats

Nous avons conçu deux plateformes logicielles tournant sur du matériel différent et supportant l'application DMSP. Nous avons aussi conçu un nouveau modèle de masquage, appelé « EBAC » (pour « Event-Based Access Control »), que nous avons implémenté sur le schéma de la base de données de l'application DMSP. Nous avons réalisé une expérimentation terrain, réalisé de nombreuses démonstrations, disséminé ces résultats auprès d'industriels et de juristes spécialisés dans la gestion de données personnelles de santé. L'application DMSP sert aussi de cas d'usage à plusieurs de nos articles scientifiques.

[29] Initiative permettant au patient de télécharger ses données médicales stockées par des centres médicaux ou des applications médicales, dans un format texte interprétable par toutes les applications estampillées « Blue Button ». Voir : http://en.wikipedia.org/wiki/The_Blue_Button

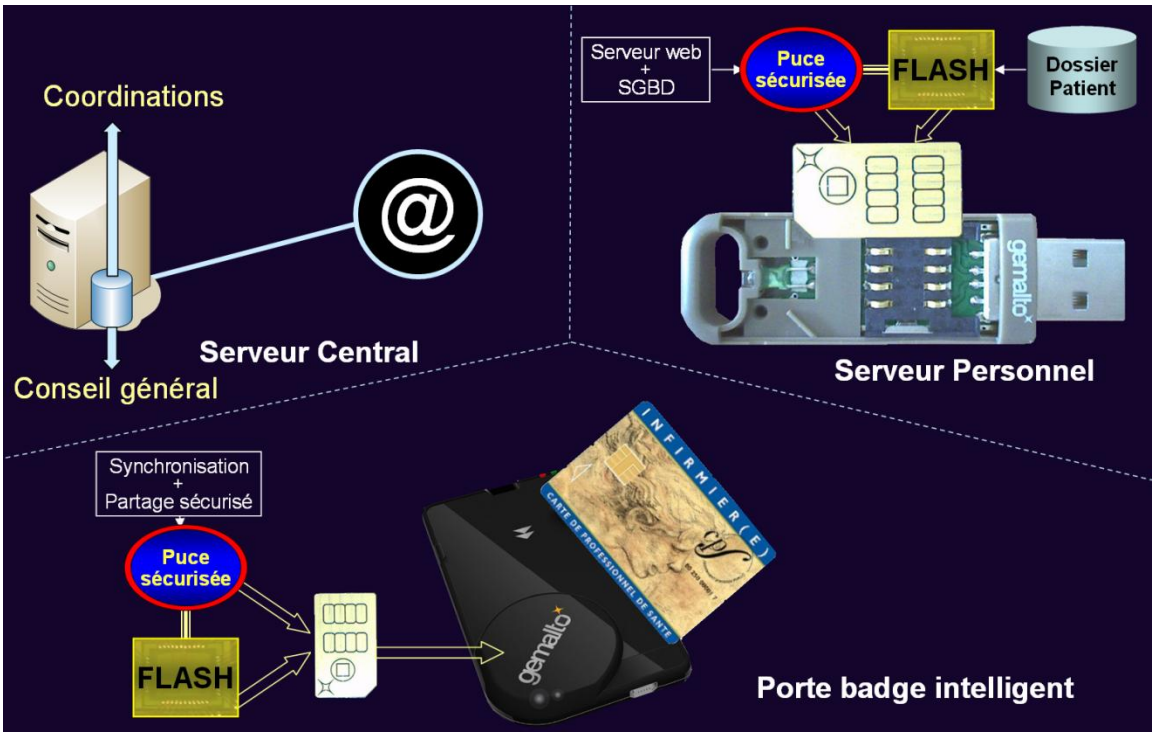


***Figure 12.*** *Architecture initiale de la plate-forme PlugDB*

***Plateforme logicielle et matérielle initiale.*** La plateforme initiale est basés sur trois éléments (voir figure 12) : (1) un Serveur Personnel (SPD) intégré dans une carte SIM de nouvelle génération sur laquelle est superposée une mémoire de type NAND Flash (256 MB) intégrée dans un châssis USB ; (2) un porte badge intelligent, lecteur de carte CPS (pour les médecins) et CPA (pour les intervenant sociaux) et qui contient aussi une carte SIM à grande mémoire permettant de convoyer les fichiers de synchronisation entre les serveurs personnels et le serveur central ; (3) un serveur central contenant une réplique (chiffrée) des données du dossier permettant d'alimenter les différentes chaînes de traitement et de régénérer le contenu du dossier en cas de perte du serveur personnel. Nos contributions logicielles à cette plateforme se sont portées sur le développement du SGBD relationnel embarqué, la conception de modules annexes permettant de pré-compiler des requêtes SQL pour les applications, les spécifications et l'implémentation du processus de synchronisation SPD/Serveur central, la conception du pilote JDBC embarqué (il n'y a pas de standard JDBC pour Javacard, le sous-ensemble de Java pour l'embarqué), et son optimisation pour le système d'exploitation CIGALE (OS Java natif, JVM embarquée) de Gemalto. Nous avons développé cette plateforme avec les ingénieurs, doctorants

et chercheurs de l'équipe SMIS de 2007 à 2010 et en interaction avec des ingénieurs de chez Santeos (Amaury Willemand et Solène Martin) et de chez Gemalto (Jean-Jacques Vandewalle, Laurent Lagosanto, Olivier Potonnier, Patrick Enjolras et Laurent Boulard).

***Expérimentation terrain.*** Sur la base de cette plateforme, nous avons conduit une expérimentation terrain sur le territoire des Yvelines, pendant 18 mois (de mi-2011 à fin 2012) auprès de 40 patients et 80 professionnels. Les patients et professionnels ont été équipés de mini-PC (modèles eePC) capables de jouer l'application DMSP depuis un navigateur Web, s'interfaçant indifféremment avec le serveur central ou le SPD. Un bilan de l'usage de l'application a été réalisé par Clarisse Chalard, étudiante en Sciences Sociales à l'UVSQ. Les conclusions étaient prometteuses (amélioration de la communication entre praticiens et entre métiers, évolution des pratiques de soins, suppression de la double-saisie) mais se heurtaient à une inadéquation de certains éléments de la technologie (trop d'éléments matériels différents à connecter au domicile, obligation d'installer des pilotes spécifiques Gemalto sur les eePC compliquant le déploiement et ayant généré des incidents techniques répétés).

***Plateforme logicielle et matérielle actuelle.*** Nous avons élaboré une deuxième plateforme, corrigeant les défauts de la première. D'abord, nous avons remplacé le SPD de Gemalto par un composant matériel de fonctionnalité équivalente, réalisable par toute PME spécialisée en électronique. Le composant a été conçu sous notre direction par la société Zed Electronics[30]. L'objectif était de maîtriser l'OS (nous utilisons maintenant FreeRTOS[31]) et la technologie hardware pour être en mesure de la faire évoluer en fonction des besoins de l'application plutôt que d'utiliser la solution fermée de Gemalto. Nous avons porté notre moteur de gestion de données sur ce nouveau matériel, ce qui nous a conduits à modifier les structures de données (buffers, journaux transactionnels) stockées en mémoire NOR (nouvelle technologie de NOR), à s'interfacer avec un nouvel OS et à intégrer des fichiers pilotes adaptés à nos périphériques. Nous avons ainsi pu rendre le SPD complètement Plug-and-Play sur tablettes et smartphone

[30] ZED est une société d'électronique spécialisée dans la conception de prototypes et dans la fabrication de petites séries. Voir http://www.zeus.fr/

[31] Système d'exploitation embarqué et open source pour microcontrôleur le plus utilisé du marché. Voir http://www.freertos.org/

Android. Nous avons aussi doté le SPD d'un module Bluetooth et des primitives nécessaires à un usage sans fil. Nous y avons intégré un lecteur d'empreinte digitale afin de s'affranchir de l'usage des cartes CPS/CPA tout en permettant une authentification forte des professionnels. Un microphone pourrait aussi être intégré sans impact sur le coût du dispositif.

**EBAC, un modèle de droits d'accès basé sur des évènements sémantiques.** Cette étude adresse le problème du recueil du consentement d'un usager sur une politique d'accès, dans le domaine de la santé. Les données de santé sont régies par des politiques d'accès si complexes qu'il est inenvisageable d'obtenir un consentement éclairé des patients comme cela est requis par la loi. La difficulté réside à la fois dans le très grand nombre de règles mises en œuvre et dans la complexité intrinsèque des données médicales. Nous avons proposé le modèle EBAC (« *Event-Based Access Control* ») pour aider le patient à réguler les données sensibles de son dossier en partant d'une sémantique connue. Ainsi les données sont regroupées sous forme d'épisodes ou d'événements (« avortement », « dépression nerveuse 2012 »). Le patient choisit les intervenants médicaux qui peuvent ou non jouer un rôle sur ces événements, et précise les modalités d'échange d'informations entre ces intervenants. EBAC est appliqué comme un modèle de masquage dont les règles ont priorité sur celles définies dans la politique de contrôle d'accès générale. Les bases du modèle EBAC ont été publiées dans des revues et chapitres de livres [AAB+09, AAB+10a, AAB+10b]. Une version simplifiée du modèle a été utilisée pour l'expérimentation terrain de l'application DMSP.

***Articles scientifiques.*** Outre les publications [AAB+09, AAB+10a, AAB+10b] liées au modèle EBAC, certaines publications scientifiques décrivent l'application DMSP [ABB+08a, ABB+08b].

***Dissémination.*** La plateforme initiale a été démontrée dans une douzaine d'événements nationaux et internationaux dont Javaone [AnV09] (15000 participants) et Futur en Seine [Anc13a]. Nos solutions ont fait l'objet d'une audition publique à l'Assemblée Nationale en 2009 de Philippe Pucheral, responsable du projet SMIS, dans le cadre de la relance du dossier médical personnel (DMP) national. L'expérimentation terrain conduite dans les Yvelines a fait l'objet de nombreuses brèves publiées par le Conseil Général. De plus le cas d'usage DMSP a servi de base à de nombreuses discussions entre juristes et informaticiens lors du projet multi

disciplinaire DEMOTIS, visant à étudier les compromis techniques et juridiques sous-jacents à la conception des infrastructures en charge du Dossier Médical Personnalisé (DMP).

## 5. Conclusion

L'application DMSP est très appréciée des professionnels de santé et des acteurs sociaux. Elle répond bien à leurs besoins et présente des garanties de sécurité inégalées permettant de garantir le respect de la vie privée des patients. La solution est implantable sur plateforme générique et n'importe quel assembleur de composant électronique peut la fabriquer pour quelques dizaines d'euros. L'application arrive à un niveau de maturité qui nous permet d'envisager une industrialisation. Un audit de la solution est actuellement conduit par l'ARS île de France afin d'étudier la possibilité d'une expérimentation plus large, préfigurant une industrialisation. Les enjeux du domaine de la santé sont particulièrement complexes et l'avenir de DMSP dépend de nombreux facteurs que nous ne maîtrisons pas. Cependant, dans une réponse[32] publiée dans le journal Officiel du Sénat daté du 21/08/2014, Marisol Touraine, Ministre de la Santé, constate que seuls 473 493 dossiers DMP ont été créés en France, la plupart étant vides ou ne contenant qu'un seul document, et a décidé de « *recentrer le DMP, renommé dossier médical partagé, sur les patients atteints de maladies chroniques ainsi que sur les personnes âgées, en particulier dans le cadre des expérimentations personnes âgées en risque de perte d'autonomie (PAERPA), qui justifient prioritairement d'une prise en charge pluriprofessionnelle coordonnée* ». Cette décision, nous l'espérons, nous aidera à envisager des expérimentations plus larges, dans lequel l'outil DMSP pourrait être vu comme facilitateur dans l'alimentation du DMP.

Enfin, l'expérience accumulée dans le cadre de DMSP nous a permis de mettre au point un nouveau composant dont nous maîtrisons le matériel et le système d'exploitation. Saliha Lalalli et Cuong To, doctorants dans l'équipe, réalisent actuellement leurs développements et validations sur ce composant. Athanasia Katsouraki, elle aussi doctorante dans l'équipe, monte actuellement une expérimentation sur l'acceptabilité de ce type de composant et sur l'impact en

[32] http://www.senat.fr/basile/visio.do?id=qSEQ120901761&idtable=q289377|q259785|q263126|q258878&_c=%22paerpa%22&rch=qs&de=20110826&au=20140826&dp=3+ans&radio=dp&aff=sep&tri=p&off=0&afd=ppr&afd=ppl&afd=pjl&afd=cvn

termes de perception de sécurité des utilisateurs, en collaboration avec des chercheurs en économie expérimentale du laboratoire ADIS dans le cadre du projet ISN. Tout ceci découle en grande partie de notre implication collective dans l'élaboration de cette application.

# Conclusion et perspectives

*Cette section conclue le document et présente certains des éléments de mon programme de recherche, les plus en ligne avec les travaux présentés dans les chapitres précédents.*

## 1. Conclusion générale

Les approches centralisées de gestion des données personnelles posent vis-à-vis du respect de la vie privée des problèmes intrinsèques, liés à un très faible ratio coût/bénéfice des attaques perpétrées et au modèle sous-jacent basé sur une délégation de la gestion des données, ce qui conduit à des usages hors de contrôle du propriétaire des données, voire indésirables.

Dans ce document, nous avons présenté différentes architectures, alternatives ou complémentaires au modèle du Web actuel qui dans différents contextes permettent de garantir une gestion de données plus respectueuse de la vie privée. Notre approche repose sur l'introduction d'un Serveur Personnel de Données (SPD) sécurisé matériellement, permettant à l'individu d'exercer un contrôle sur ses données personnelles, leur partage et leur dissémination, avec un niveau de garantie inégalé, hérité de la sécurisé physique offerte par le SPD. Ces différentes architectures permettent d'envisager l'émergence d'un nouveau modèle que nous appelons le « Web Personnel Sécurisé ».

L'étude de ces architectures nous a conduits à proposer différentes contributions scientifiques relatives à leur mise en œuvre. Deux d'entre elles sont présentées plus en détail dans ce manuscrit. La première concerne la conception d'un SGBD embarqué compatible avec les fortes contraintes du dispositif. Le design du SGBD se base sur des structures d'indexation massive permettant d'évaluer efficacement les requêtes avec très peu de RAM, générées séquentiellement pour éviter les mises à jour aléatoires en mémoire NAND Flash et réorganisées de façon itérative pour supporter de grands volumes de données. La seconde contribution présentée est centrée sur le problème de la minimisation de l'exposition de données transmises à des tiers, dans le cas d'interactions avec des services en ligne au travers de formulaires. D'autres contributions parfois évoquées dans ce manuscrit n'y ont pas été présentées. Elles concernent par exemple la dégradation progressive des données personnelles [HFA09, ABH+08a, ABH+08b, ABH+08c, HAF+06], l'exécution de requêtes croisant des données privées embarquées avec des données externes publiques sans fuite d'information

privée [AAB+07, SAB+07, AAB+09] ou de nouveaux modèles permettant aux individus de réguler le partage de leurs données personnelles [AAB+10b, AAB+09, AAB+10c]. Les bénéfices potentiels de notre approche ont été illustrés au travers de l'application DMSP.

Mon programme de recherche des prochaines années vise à contribuer à l'établissement d'un « Web personnel sécurisé ». Pour disséminer cette vision, nous envisageons de proposer une version en logiciel libre et matériel libre de nos solutions permettant un usage et un développement communautaire. La dissémination de cette plateforme commence par une mise à disposition auprès des étudiants de différentes universités et écoles d'ingénieurs. Nous travaillons aussi en lien avec la startup CozyCloud qui propose une plateforme libre de « Web personnel », afin d'y intégrer de plus fortes garanties grâce à nos technologies.

La suite de ce chapitre présente certaines perspectives de recherches, les plus en lien avec notre vision du « Web personnel sécurisé ». Ces perspectives sont structurées en trois axes : (1) l'embarquement de nouvelles fonctionnalités de gestion de données embarqué sous-jacentes au partage sécurisé des données, (2) l'étude de modèles de contrôle d'usage permettant de sécuriser les traitements effectués sur les données personnelles et qui ne peuvent être embarqués et (3) la gestion de données personnelles en l'absence d'infrastructure. Ces trois axes sont nécessaires et complémentaires en vue de la réalisation d'un « Web personnel sécurisé ». Le troisième axe étant plus particulièrement pertinent dans le cas des Pays les Moins Avancés.

## 2. Gestion de données embarquées pour sécuriser les contrôles d'accès

Dans le contexte du Web personnel sécurisé, garantir le partage des données suivant des règles d'accès établies nécessite d'évaluer dans le composant sécurisé certaines opérations de gestion de données. Par exemple, pour une base de données relationnelle, c'est le calcul des vues autorisées qui doit être embarqué. Pour d'autres modèles de données (documents, clé valeur, etc.) et d'autres modèles de contrôle d'accès (expressions portant sur des tags, modèles de contrôle d'accès basés attributs, etc.), il est nécessaire de pouvoir embarquer d'autres techniques de gestion de données. Nous envisageons de définir des méthodes de conceptions permettant une transposition systématique dans l'environnement embarqué des techniques de gestion de données sous-jacentes à l'évaluation de règles d'accès. Nous illustrons ici cette problématique sur le cas de la gestion de collections de documents accessibles au travers de fonctions de recherche d'information sur le contenu.

**Motivation.** Dans le contexte du Web personnel sécurisé, les documents relatifs à un individu (mails, fichiers divers, administratifs, médicaux, bancaires, etc.) sont régulés depuis le serveur personnel. Des droits sont donnés aux applications en fonction des tags apposés sur les fichiers. Certaines applications (gestion des mails, gestion de fichiers, etc.) nécessitent un accès aux fichiers au travers d'un moteur de recherche d'information. Le moteur de recherche doit être embarqué pour éviter d'exposer hors de l'enceinte sécurisée l'index inversé permettant de réaliser les recherches (il contient des informations sur le contenu de chaque document indexé). Lors d'une recherche seuls les documents qualifiés par la recherche et compatibles avec les règles d'accès établies doivent être produits en résultat.

Les techniques de recherche d'information pour l'embarqué se justifient aussi dans d'autres contextes. En effet de nombreux objets intelligents de notre environnement sont maintenant dotés de capacités (locales) de collecte, de stockage et de traitements de données. Ils offrent des interfaces de recherche d'information pour accéder aux données [YGM08]. Certains objets domotiques maintiennent une description de leur environnement [YSM+08]. Par exemple, certaines librairies offrent la possibilité aux lecteurs d'interroger directement les rayons pour retrouver les livres les plus pertinents. D'autres exemples justifient la conception de moteurs de recherche embarqués dans des capteurs pour retrouver les objets pertinents de leur environnement ou dans des appareils photo (ou capteurs photos) pour rechercher des images sur des tags.

Tous cela justifie la conception d'un moteur de recherche pouvant être vu comme une généralisation pour l'embarqué de Google Desktop ou de Spotlight. Dans le contexte du Web personnel sécurisé ce moteur devra être capable de traiter de grandes collections de documents et d'évaluer des droits d'accès basés sur des tags associés à ces documents.

***Etat de l'art.*** Les moteurs de recherche ont été largement étudiés dans la littérature (voir [ZoM06] pour un état de l'art). Ils se basent sur des index inversés et sur une fonction de classement (par exemple basée sur une métrique de type *tf-idf* [33]) pour retrouver les documents les plus pertinents pour une recherche. Chaque document est associé à un ensemble de termes (décrivant le contenu du document) pondérés (caractérisant l'importance du terme dans le

[33] « Term Frequency-Inverse Document Frequency ». Métrique traditionnelle en recherche d'information permettant de classer des documents selon leur pertinence, en fonction de la fréquence des termes de la recherche dans le document, et de l'inverse de la fréquence de ces termes dans la collection complète.

document). Pour les documents textuels, les termes sont les mots composant le document et leur poids est la fréquence du mot dans le document. Un index inversé, organisé sous forme d'arbre, associe à chaque terme la liste des documents contenant ce terme et le poids du terme dans le document. Les requêtes sont évaluées de la manière suivante : (1) l'index inversé est accédé pour chaque terme de la requête et renvoie les listes de couples (identifiant du document, poids) correspondantes, (2) un container est alloué en RAM pour chaque identifiant de document retourné par l'index et stocke les poids correspondants pour chaque terme de la recherche, (3) le score *tf-idf* est calculé pour chaque document/container, puis (4) les documents sont triés par score et les *k* plus grands sont produits en résultat. Cette stratégie ne peut pas être appliquée dans le contexte embarqué car d'une part l'index inversé ne peut pas être maintenu en Flash NAND efficacement sous forme arborescente (le temps d'insertion de nouveaux documents serait inacceptable) et d'autre part la quantité de RAM disponible est très insuffisante pour permettre l'allocation d'un container par document.

Certains travaux de recherche se sont intéressés à concevoir des techniques de recherche d'information pour l'embarqué [TSW+08, TSW+10, WTL10, YGM08]. Elles sont basées sur une organisation séquentielle de l'index inversé mais ne fonctionnent qu'avec un petit nombre de documents (quelques centaines) et ne passent pas à l'échelle (la performance des recherches est linéaire avec le nombre de documents indexés). D'autre part elles ne permettent pas de supporter la suppression de documents.

***Approche.*** La difficulté du problème vient des contraintes conflictuelles entre une RAM très limitée et une mémoire Flash NAND supportant très mal les petites mises à jour aléatoires. Notre approche se fonde sur trois principes de conception: (1) l'index inversé doit être composé de partitions écrites séquentiellement et jamais modifiées, (2) les requêtes doivent être évaluées sur les différentes partitions dans une RAM bornée avec un coût du même ordre de magnitude que le coût équivalent sur une structure optimale (ex. un index inversé sous forme d'arbre) et (3) pour limiter le nombre total de partitions celles-ci doivent être fusionnées en RAM bornée sans générer (trop) d'écritures aléatoires. Nous cherchons à définir à partir de ces principes un nouvel index inversé adapté à la recherche d'information sur de grands volumes de données (centaines de milliers de documents) et implantant des contrôles d'accès. Dans un deuxième temps, nous chercherons à affiner les principes de conception et à couvrir les bases de données clé/valeur et les modèles de contrôle d'accès associés. Plus généralement, nous essaierons de délimiter le champ d'application de nos principes de design de manière à permettre une transposition

systématique dans l'environnement embarqué des techniques de gestion de données sous-jacentes à l'évaluation de règles d'accès.

## 3. Contrôle d'usage

Certains traitements orientés données complexes ne peuvent être embarqués dans l'enceinte sécurisée d'un serveur personnel, car ils consomment beaucoup trop de ressources (calcul de profil, de recommandations, etc.). Mais les exécuter hors du serveur personnel nécessite d'externaliser de grands volumes de données personnelles nécessaires au traitement, ce qui peut remettre en cause la confidentialité des données. Il est donc nécessaire d'étudier et de proposer des techniques permettant de contrôler l'usage que certaines applications font des données, au risque de rendre l'approche Web personnel sécurisé caduque pour les traitements de données complexes. Cette étude consiste à étendre la sphère de sécurité, qui se limite essentiellement jusqu'ici à garantir des règles de contrôles d'accès, pour garantir du contrôle d'usage[34]. Par contrôle d'usage, nous désignons la possibilité de contrôler ce qu'une application fait des données en limitant les effets de bord de cette application, c'est-à-dire toute exploitation ou fuite de données qui pourrait conduire à un usage indésirable des données.

Nous avons choisi d'étudier ce problème dans le contexte concret et particulièrement représentatif de la gestion de données issues de compteurs électriques intelligents. Les données de consommation électriques produites à domicile nécessitent en effet des traitements complexes, par exemple de désagrégation, pour être exploitées. Nous illustrons ici la problématique du contrôle d'usage sur ce cas particulier.

***Motivation.*** Dans les prochaines années, de nombreux foyers en Europe seront équipés de compteurs intelligents. Ces compteurs produisent une trace détaillée de la consommation électrique d'un logement. L'objectif de cette technologie est d'une part de permettre aux distributeurs d'énergie de mieux observer le réseau pour faire de l'équilibrage de charge, et d'autre part de permettre l'éclosion de nouveaux services pour les usagers, basés sur l'exploitation de leurs données de consommation d'énergie (conseils conduisant à des

[34] Notons que le terme contrôle d'usage est parfois utilisé pour désigner des modèles de contrôle d'accès sophistiqués, prenant par exemple en compte le contexte de la requête (heure, type de machine cliente, adresse IP, etc.). Il n'est pas entendu ici dans ce sens.

économies d'énergie, diagnostic en cas de panne d'un appareil électrique, jeux conduisant à faire baisser sa consommation, etc.). Le risque d'atteinte à la vie privée est réel puisque la signature énergétique révèle l'activité précise des habitants du logement (ex. [MWB11]). La CNIL Européenne signale aussi ces dangers[35]. Pour pouvoir utiliser ces données tout en limitant les risques, il faut offrir aux développeurs d'applications et aux usagers une plateforme permettant de fournir des garanties quant à l'usage effectif que les applications font des données.

***Etat de l'art.*** L'usage par le propriétaire des données issues de son compteur intelligent n'est pas régulé et permet à tout citoyen Européen d'utiliser des applications en ligne offrant un certains services sur la base de ces données. Aux Etats-Unis l'initiative « Green Button[36] », à l'image du « Blue Button » pour les données de santé, permet aux individus de télécharger leurs données de consommation d'énergie auprès de leur fournisseur dans un format textuel compréhensible et documenté, utilisable par de nombreuses applications en ligne (notamment toutes les applications estampillées « Green Button »). Une telle exploitation des données n'offre pas de contrôle à l'usager sur la façon dont ses données sont utilisées ou disséminées.

Certaines solutions de l'état de l'art prennent en compte le problème du respect de la vie privée pour les données issues de compteurs intelligents. Certaines propositions visent à intégrer dans le compteur des fonctions capables de facturer l'usager en supportant différentes politiques tarifaires sans externaliser les données de consommation électrique. Notamment des protocoles sécurisés ont été proposés pour couvrir les usages principaux du fournisseur d'électricité, sans exposer les données du compteur, tout en ayant une preuve d'authenticité du résultat [RiD11]. D'autres techniques permettent de réaliser des calculs d'agrégats provenant d'un ensemble de compteurs en se basant sur des techniques de chiffrement homomorphe (ex. [LLL10]). Ces travaux s'intéressent uniquement à des usages spécifiques des données de compteurs,

[35] Voir le communiqué de presse du 11 juin 2012 du Contrôleur européen de la protection de données (EDPS), intitulé « Compteurs intelligents: le profilage des consommateurs permettra de suivre bien plus que leur consommation d'énergie si des limites claires ne sont pas établies ». Extrait du communiqué: l'introduction d'un compteur intelligent permet potentiellement de « *suivre ce que les membres d'un ménage font dans l'intimité de leurs maisons, s'ils sont en vacances ou au travail, si l'un d'eux utilise un dispositif médical spécifique ou un moniteur pour bébé, comment ils aiment passer leur temps libre, etc.* »

[36] Voir : http://www.nist.gov/smartgrid/greenbutton.cfm

nécessaires aux autorités publiques ou aux organismes intervenant dans la distribution d'électricité et dans la facturation. Mais ils ne sont pas adaptés aux nouvelles applications personnelles possibles, en plein essor, permettant à l'usager d'analyser et de partager ses données, et d'améliorer sa propre consommation énergétique.

***Approche.*** Les opérations à appliquer à la trace de consommation électrique d'un usager sont consommatrices de ressources et brassent de très grands volumes de données (potentiellement toute la trace). C'est le cas de l'opération de désagrégation qui apporte une sémantique aux données de consommation (ex. une vue des appareils électriques en fonctionnement). La désagrégation nécessite d'identifier la signature électrique de chaque appareil dans la trace, calcul qui ne peut pas être évalué dans l'environnement contraint du serveur embarqué. Il est pourtant nécessaire de supporter ce type d'opération qui précède à tout usage applicatif des données. Notre objectif est donc de déporter ce type de calcul hors de l'enceinte sécurisée, tout en garantissant à l'individu que l'opération s'exécute sans effet de bord indésirable (c'est-à-dire sans fuite des données qu'elle manipule). Ce modèle de contrôle d'usage pourra ensuite être étendu aux autres applications orientées données.

Pour contrôler ces effets de bord, nous envisageons d'isoler l'environnement d'exécution de certains traitements (de certains périphériques, et des autres traitements applicatifs) grâce à des techniques de « sandboxing ». Cela nécessite d'implanter une infrastructure telle qu'illustrée sur la figure 13. Le compteur alimente un serveur personnel sécurisé capable de communiquer (en Wifi, Bluetooth ou USB) avec l'infrastructure d'exécution des applications et d'appliquer des fonctions utilisateurs orientées données trop complexes pour être embarquées.

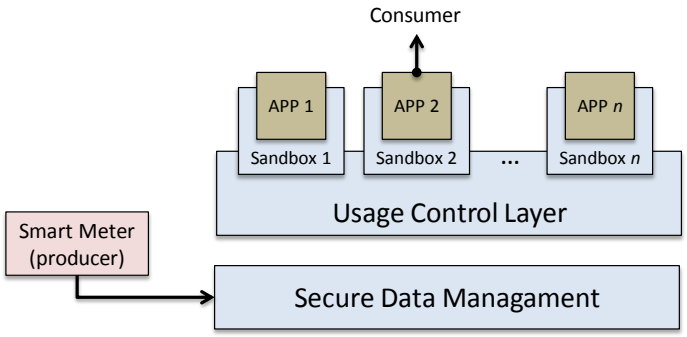


***Figure 13.*** *Contrôle d'usage basé sur le principe du bac à sable (sandboxing).*

Tout traitement orienté données externe peut occasionner une fuite d'information (effet de bord), soit via un canal de transmission auxiliaire (accès à un périphérique réseau, utilisation du système de fichier), soit via son résultat qui peut être non conforme et intégrer des informations supplémentaires. Pour éviter le premier type de fuite, nous envisageons de permettre un découpage des applications de manière à (1) isoler les traitements faisant des accès intensifs aux données (par exemple pour désagréger la trace énergétique) du reste de l'application ; (2) réguler le flux de données entre les compartiments isolés de l'application grâce au serveur personnel ; et (3) calibrer correctement les autorisations et les accès aux périphériques nécessaires à chacun des compartiments de l'application. L'isolation peut être obtenue grâce à Trustzone (voir chapitre 1, section 4.1) ou avec un hyperviseur classique (ex. Xen, pour lequel l'isolation peut être renforcée par du matériel sécurisé). Pour éviter le second type de fuite (via le résultat du calcul), des techniques de tests devront être définies (techniques collaboratives entre usagers, statistiques, à base de jeux d'entrées/sorties connues, etc.).

L'étude de cette perspective de recherche a tout juste démarré dans le cadre d'une collaboration avec Philippe Bonnet, Professeur à l'ITU au Danemark, qui a passé une année au sein du projet SMIS en 2013/2014.

## 4. Gestion de données sans infrastructure

D'après de nombreuses études [ITU11, CBP+12], les technologies de l'information constituent pour les Pays les Moins Avancés un élément facilitateur de développement économique pouvant améliorer la situation dans des domaines importants comme l'éducation, la santé ou la finance. Nous pensons qu'un serveur personnel de données permettrait d'offrir des services orientés données dans certains de ces différents domaines. Cependant, le contexte des Pays les Moins Avancés nous conduit à considérer une approche sans infrastructure (voir chapitre 1, section 4.3), très redondante et robuste par construction, ne reposant sur aucune autorité centrale pour fonctionner, où les communications sont asynchrones et opportunistes et où les coûts globaux du système sont directement proportionnels à la taille de la population visée. Les problèmes à résoudre pour établir un serveur personnel dans ce contexte concernent par exemple le déploiement d'applications, la modélisation de données unifiées, la vérification des identités, les règles de partage des données, les requêtes distribuées, tout cela sans reposer

sur aucune infrastructure centrale et avec une gestion opportuniste des communications. Le problème de l'authentification et du contrôle d'accès sans autorité centrale est détaillé ci-dessous à titre d'illustration.

***Motivation.*** Les techniques de vérification des identités se basent sur des infrastructures à clé publique (PKI). Elles sont généralement centralisées, et reposent sur l'existence d'autorités de certification, qui sont les seules à être de confiance dans l'infrastructure. La vérification des identités s'organise alors sous forme hiérarchique à partir de ces autorités. Dans un contexte où le tissu économique, politique, associatif, la société civile, sont fortement représentés et bien structurés, les autorités de confiance peuvent être établies par consensus. C'est indirectement sur l'existence de toute cette structuration que repose la confiance. Nous considérons que ce modèle n'est pas transposable dans le contexte des PMA, qui se caractérisent par l'absence de structuration.

***Etat de l'art.*** Certaines techniques de vérification des identités, alternatives et décentralisées, ont été proposées, comme PGP. Ces techniques reposent sur la notion de toile de confiance. Dans les approches basées sur le standard OpenPGP, chaque individu détient une clé publique, une clé privée et un trousseau de clés permettant de stocker des clés et des signatures, et peut certifier l'identité d'un autre individu. Un individu A peut alors certifier l'identité d'un individu B par signature (A signe la clé publique de B avec sa clé privée). La vérification des identités repose sur le niveau de confiance que chacun attribue aux autres individus. Ainsi, si Alice fait confiance à Bob (dans le fait qu'il identifie correctement les individus qu'il certifie), les clés signées par Bob pourront être considérées comme valides par Alice. Des niveaux de confiance gradués peuvent aussi être attribués aux utilisateurs, et la validité d'une identité pourra alors être obtenue si suffisamment d'individus de niveau de confiance intermédiaire ont certifié cette identité. Ces techniques constituent un point de départ intéressant mais elles ne peuvent pas être transposées directement dans notre contexte sans infrastructure. En effet des serveurs sont nécessaires pour stocker et distribuer les clés et les signatures. L'approche suppose aussi qu'un nombre d'individus suffisamment important participe au système (des « signing parties » doivent être organisées au départ pour que les individus s'authentifient les uns les autres). Enfin cette approche ne résout pas le problème de la propagation de règles d'accès.

***Approche.*** Une caractéristique sous-jacente de l'architecture Folk-IS est de déporter les outils de contrôle aux extrémités du réseau dans du matériel sécurisé et de tirer partie des interactions en face-à-face entre les individus. Cela peut permettre d'établir des sphères de confiance locales. Par exemple, une organisation locale (ex. une ONG de suivi sanitaire) pourrait produire des applications, un modèle de données, des informations d'identifications ad-hoc et des politiques de contrôle d'accès, et les pousser sur le réseau Folk-IS (l'injection des données pourrait être conduite localement par des interactions physiques entre des individus et les acteurs de terrain de l'ONG). Nous espérons ainsi pouvoir jeter les bases d'un nouveau mode de gestion de données, semi-centralisé, dans lequel chaque *Folk-node* pourrait garantir (1) que les données produites par les acteurs d'une organisation donnée ne peuvent fuiter hors de l'organisation, et (2) que la vie privée des individus participant au système est toujours respectée.

Nous sommes actuellement en train de monter une proposition de projet européen répondant à l'appel H2020 ICT39[37], en collaboration avec notamment l'équipe IDASCO du LIRIMA au Cameroun.

[37] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/9094-ict-39-2015.html

# Bibliographie


[AAB+09] T. Allard, N. Anciaux, L. Bouganim, P. Pucheral, R. Thion. Seamless Access to Healthcare Folders with Strong Privacy Guarantees. Special issue of the Journal of Healthcare Delivery Reform Initiatives, Vol. 1, n°4, pp. 82-107, 2009.

[AAB+10a] T. Allard, N. Anciaux, L. Bouganim, Y. Guo, L. Le Folgoc, B. Nguyen, Pucheral P. , Ray I. , Ray I., and Yin S. Secure personal data servers: a vision paper. 36th International Conference on Very Large Data Bases (VLDB), pp. 25-35, 2010.

[AAB+10b] T. Allard, N. Anciaux, L. Bouganim, P. Pucheral, R. Thion. Chap. IX: Trustworthiness of Pervasive Healthcare Folders. Book chapter of Pervasive and Smart Technologies for Healthcare: Ubiquitous Methodologies and Tools, A. Coronato, G. De Pietro (editors), Information Science Reference, pp. 1-24, 2010.

[AAB+10c] T. Allard, N. Anciaux, L. Bouganim, P. Pucheral, R. Thion. Concilier ubiquité et sécurité des données médicales. Les Cahiers du CRID « Les technologies au service des droits, opportunités, défis, limites », D. Le Métayer (Editor), Vol. 32, Jan. 2010.

[AAG+98] Alimonti, P., Ausiello, G., Giovaniello, L., , and Protasi, M. On the Complexity of approximating weighted satis_ability problems. Tech. rep., Universit_a di Roma, 1998.

[ABB+07] N. Anciaux, M. Benzine, L. Bouganim, P. Pucheral, D. Shasha: GhostDB: Querying Visible and Hidden Data without Leaks. 26th ACM International Conference on Management of Data (ACM SIGMOD), Beijing, China, June 2007.

[ABB+08a] N. Anciaux, M. Berthelot, L. Braconnier, L. Bouganim, M. De la Blache, G. Gardarin, P. Kesmarszky, S. Lartigue, J-F. Navarre, P. Pucheral, J-J. Vandewalle, K. Zeitouni. A Tamper-Resistant and Portable Healthcare Folder. International Journal of Telemedicine and Applications (IJTA), Vol. 2008, 9 pages, 2008.

[ABB+08b] N. Anciaux, M. Benzine, L. Bouganim, K. Jacquemin, P. Pucheral, S. Yin. Restoring the Patient Control over her Medical History. 21th IEEE Int. Symposium on Computer-Based Medical Systems (IEEE CBMS), Jyväskylä, Finland, June 2008.

[ABB+09] N. Anciaux, M. Benzine, L. Bouganim, P. Pucheral, D. Shasha. Revelation on Demand. Distributed and Parallel Database Journal (DAPD), Vol. 25, n°1-2, pp. 5-28, 2009.

[ABB+13] N. Anciaux, P. Bonnet, L. Bouganim, B. Nguyen, P.Pucheral, I. S. Popa. Trusted Cells : A Sea Change for Personal Data Services. 6th Conference on Innovative Database Research (CIDR), 4 p., 2013.

[ABB+14] N. Anciaux, P. Bonnet, L. Bouganim, P. Pucheral. Trusted Cells: Ensuring Privacy for the Citizens of Smart Cities. ERCIM News, Vol. 98, 2014.

[ABD+13] N. Anciaux, L. Bouganim, T. Delot, S. Ilarri, L. Kloul, N. Mitton, P. Pucheral. Folk-IS: Opportunistic Data Services in Least Developed Countries. 36th International Conference on Very Large Data Bases (PVLDB), Vol. 7(5), Vision Paper, pp. 425-428, 2014.

[ABD+14] N. Anciaux, L. Bouganim, T. Delot, S. Ilarri, L. Kloul, N. Mitton, P. Pucheral. Opportunistic data services in least developed countries: benefits, challenges and feasibility issues. SIGMOD Record, Vol. 43, n°1, pp. 52-63, 2014.

[ABG+10] N. Anciaux, L. Bouganim, Y. Guo, P. Pucheral, J.-J. Vandewalle, S. Yin. Pluggable Personal Data Servers. 29th ACM International Conference on Management of Data (ACM SIGMOD), demo. Paper, Indianapolis, Indiana, Jun. 2010.

[ABH+08a] N. Anciaux, L. Bouganim, H. van Heerde, P. Pucheral, P. M. G. Apers. Data Degradation: Making Private Data Less Sensitive Over Time. 17th ACM International Conference on Information and Knowledge Management (ACM CIKM), short paper, Napa Valley, USA, to appear, Oct. 2008.

[ABH+08b] N. Anciaux, L. Bouganim, H. van Heerde, P. Pucheral, P. M. G. Apers. InstantDB: Enforcing Timely Degradation of Sensitive Data. 24th International Conference on Data Engineering (ICDE), short paper, Cancun, Mexico, Apr. 2008.

[ABH+08c] N. Anciaux, L. Bouganim, H. van Heerde, P. Pucheral, P. M. G. Apers. Dégradation progressive et irréversible des données. 24èmes journées Bases de Données Avancées (BDA), Oct. 2008.

[ABN+12] N. Anciaux, D. Boutara, B. Nguyen, M; Vazirgiannis. Limiting Data Exposure in Multi-Label Classification Processes. In International Workshop on Privacy-AwaRe Intelligent Systems (PARIS2012), 2012.

[ABN+13] N. Anciaux, W. Bezza, B. Nguyen, M. Vazirgiannis. MinExp-Card : Limiting Data Collection Using a Smart Card. 16th International Conference on Extending Database Technology (EDBT), demo paper, pp. 753-756, 2013.

[ABN+15] N. Anciaux, D. Boutara, B. Nguyen, M. Vazirgiannis. Limiting Data Exposure in Multi-Label Classification Processes. Fundamenta Informaticae, to appear in 2015.

[ABP+01] N. Anciaux, C. Bobineau, L. Bouganim, P. Pucheral, P. Valduriez, ''PicoDBMS: Validation and Experience. 27th International Conference on Very Large Data Bases (VLDB), demo. paper, Roma, September 2001.

[ABP+07] N. Anciaux, L. Bouganim, P. Pucheral, 'Future Trends in Secure Chip Data Management', IEEE Data Engineering Bulletin (IEEE DEB), Vol. 30, n°3, 2007.

[ABP+08a] N. Anciaux, L. Bouganim, P. Pucheral, K. Jacquemin, S. Yin, D. Shasha, C. Salperwyck, M. Benzine. Software: PlugDB-engine version 1, registered at the 'Agence pour la Protection des Programmes (APP)' under the reference IDDN.FR.001.280004.000.S.C.2008.0000.10000, July 2008.

[ABP+08b] N. Anciaux, L. Bouganim, P. Pucheral, P. Valduriez. DiSC: Benchmarking Secure Chip DBMS. IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), Vol. 20, n° 10, pp. 1363-1377, 2008.

[ABP+08c] N. Anciaux, L. Bouganim, P. Pucheral. SGBD embarqué dans une puce : retour d'expérience. Technique et Science Informatiques (TSI), Vol. 27, n°1-2, 2008.

[ABP+09] N. Anciaux, L. Bouganim, P. Pucheral, S. Yin, M. Benzine, K. Jacquemin, D. Shasha, C. Salperwyck, M. El Kholy. Software: PlugDB-engine version 2, registered at the 'Agence pour la Protection des Programmes (APP)' under the reference IDDN.FR.001.280004.000.S.C.2008.0000.10000, April 2009.

[ABP+11] N. Anciaux, L. Bouganim, P. Pucheral, S. Yin, Yanli Guo, K. Jacquemin. Software: PlugDB-engine version 3, registered at the 'Agence pour la Protection des Programmes (APP)' under the reference IDDN.FR.001.280004.000.S.C.2008.0000.10000, Nov. 2011.

[ABP+14] N. Anciaux, L. Bouganim, P. Pucheral, Y. Guo, L. Le Folgoc. MiloDB: a Personal, Secure and Portable Database Machine. Distributed and Parallel Databases (DAPD), Vol. 32, n°1, pp. 37-63, 2014.

[ABP03a] N. Anciaux, L. Bouganim, P. Pucheral. Database Components on Chip. ERCIM News, Vol. 54, July 2003.

[ABP03b] N. Anciaux, L. Bouganim, P. Pucheral. Memory Requirements for Query Execution in Highly Constrained Devices. 29th International Conference on Very Large Data Bases (VLDB), Berlin, September 2003.

[ABP06] N. Anciaux, L. Bouganim, P. Pucheral. Data confidentiality: to which extent cryptography and secured hardware can help. Annals of telecom, Vol. 61, n°3-4, 2006.

[ABP09] N. Anciaux, L. Bouganim, P. Pucheral. A Hardware Approach for Trusted Access and Usage Control. Book chapter of the Handbook of Research on Secure Multimedia Distribution, S. Lian, Y. Zhang (editors), Information Science Reference, pp. 157-179, 2009.

[ADF+12] Ardagna, C.A., De Capitani di Vimercati, S., Foresti, S., Paraboschi, S., and Samarati, P. Minimising Disclosure of Client Information in Credential-Based Interactions. Int. Journal of Information Privacy, Security and Integrity, 1(2), pp. 205-233, 2012.

[AGP10] Sécurité des bases de données. N. Anciaux, D. Gross-Amblard, P. Pucheral, R. Thion. Ecole de printemps « MASSES DE DONNEES DISTRIBUEES », Ecole de Physique de Houches, du 16 au 21 mai 2010.

[AGS+09] Agrawal, D., Ganesan, D., Sitaraman, R., Diao, Y. and Singh, S. Lazy-adaptive tree: An optimized index structure for flash devices. Proc. of the VLDB, 2(1):361-372, 2009.

[AHK+03] Ashley, P., Hada, S., Karjoth, G., Powers, C., and Schunter, M. Enterprise privacy authorization language 1.2 (EPAL 1.2). W3C Member Submission, 2003.

[AKS+02] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu: Hippocratic Databases. International Conference on Very Large Data Bases (VLDB), pp. 143-154, 2002.

[AnB06] N. Anciaux, L. Bouganim. Data Management in Embedded Smart Devices. Tutorial donné à la Smart University, co-organisée avec la 7ème édition de la conférence internationale e-smart. Sept. 2006.

[AnB14] N. Anciaux, B. Nguyen. Limiter la collecte des données personnelles, un problème juridique NP-difficile. Magazine Tangente, numéro hors-série n°52 bibliothèque, Mathématiques & Informatique, 2014.

[Anc04] Thèse de doctorat de l'Université de Versailles St-Quentin-en-Yvelines, '*Systèmes de gestion de base de données embarqués dans une puce électronique*'. Déc. 2004. Rapporteurs: Patrick Valduriez (Directeur de Recherche, Inria), Michael J. Franklin (Professeur, Univ. Berkeley). Examinateurs : Philippe Bonnet (Professeur, DIKU), Jean-Claude Marchetaux (Ingénieur de Recherche, Gemalto). Directeur : Philippe Pucheral (Professeur, UVSQ). Co-encadrant : Luc Bouganim (Directeur de Recherche, Inira).

[Anc10] N. Anciaux. Dossier Médico-Social Portable et Sécurisé. Présentation et démonstration. Les Industries du Numérique pour la Santé, RII, in conjunction with the Connectathon, 2010, Cité Mondiale, Bordeaux.

[Anc11] N. Anciaux. Dossier Médico-Social Portable et Sécurisé. Présentation et démonstration. Les sciences du numérique au service de la santé à domicile et de l'autonomie, RII, 2010, Espaces Cap 15, Paris.

[Anc13a] N. Anciaux. Gestion de données personelles respetueuse de la vie privée. Presentation et démonstration, Futur en Seine, Archipelle des projets, 2013.

[Anc13b] N. Anciaux. Une nouvelle approche de la protection de nos données. Interview, MyScienceWork news, par Abby Tabor, 2013.

[Anc14] N. Anciaux. Vers un modèle de gestion des données respectueux de la vie privée : application à la collecte limitée d'informations personnelles. Seminaire IREP "BIG DATA", 2014.

[Anc15] N. Anciaux. Garantir la confidentialité des données personnelles. Futur en Seine 2014, Répondre aux défis des smart cities, 2014.

[And04] Anderson, A.H. An Introduction to the Web Services Policy Language (WSPL). In Proceedings of the POLICY Workshop, 2004.

[ANP13] N. Anciaux, B. Nguyen, I. S. Popa. Personal Data Management with Secure Hardware : The advantage of Keeping you Data at Hand. 14th International Conference on Mobile Data Management (MDM), Advanced Seminar, pp 1-2, 2013.

[ANP14] N. Anciaux, B. Nguyen, I. S. Popa. Tutorial: Managing Personal Data with Strong Privacy Guarantees. 17th International Conference on Extending Database Technology (EDBT), Tutorial, pp. 672-673, 2014.

[AnV09] Demonstration of electronic Health Records (eHR) on Java Card™ 3.0 Technology. Nicolas Anciaux (Inria) and Jean-Jacques Vandewalle (Gemalto). BOF-4576, CS Advanced Based Devices, JavaOne Conference, San Francisco, USA, Jun. 2009.

[ANV11] N. Anciaux, B. Nguyen, M. Vazirgiannis. Minimum Exposure - A New Approach for Limited Data Collection. Invited talk, Digiteo workshop on Web Mining, 2011, Telecom ParisTech, organized by M. Vazirgiannis and P. Senellart.

[ANV12a] N. Anciaux, B. Nguyen, M. Vazirgiannis. Limiting Data Collection in Application Forms : A real-case Application of a Founding Privacy Principle. 10th Conference on Privacy, Security and Trust (PST), 8p., 2012.

[ANV12b] N. Anciaux, B. Nguyen, M. Vazirgiannis. The Minimum Exposure Project: Limiting Data Collection in Online Forms. ERCIM News, Vol. 90, 2012.

[ANV13] N. Anciaux, B. Nguyen, M. Vazirgiannis M. Exposition minimum de données pour des applications à base de classifieurs. Ingénierie des Systèmes d'Information, Vol. 18, n°4, pp. 59-85, 2013.

[APP+12] N. Anciaux, J.M. Petit, P. Pucheral, K. Zeitouni. Personal Data Server: Keeping Sensitive Data under the Indiviual's Control. ERCIM News, Vol. 90, 2012.

[Arg03] Arge L., "The Buffer Tree: A Technique for Designing Batched External Data Structures", Algorithmica, 2003.

[BaS11] S. Bajaj, R. Sion: TrustedDB: a trusted hardware based database with privacy and data confidentiality. SIGMOD Conference 2011: 205-216

[BaS14] Bajaj, S., & Sion, R. (2014). TrustedDB: A Trusted Hardware-Based Database with Privacy and Data Confidentiality. Knowledge and Data Engineering, IEEE Transactions on, 26(3), 752-765.

[BBD14] G. Blank, G. Bolsover, E. Dubois. A New Privacy Paradox. Global Cyber Security Capacity Centre, Draft Working Paper, 2014.

[BLM+09] Belotti, P., Lee, J., Liberti, L., Margot, F., Wachter, A. Branching and bounds tightening techniques for non-convex MINLP. Optimization Methods and Software 24, 4-5 (2009).

[BRD11] Bernstein P., Reid C., Das S., "Hyder - A Transactional Record Manager for Shared Flash," CIDR, 2011.

[BSS+03] Bolchini C., Salice F., Schreiber F., Tanca L., "Logical and Physical Design Issues for Smart Card Databases," TOIS, 2003.

[Cas13] D. Castro, D. How much will PRISM cost the US cloud computing industry. The Information Technology and Innovation Foundation. Jan. 2013.

[CBP+12] Coceres, R., Belding, E. M., Parikh, T. S., and Subramanian, L. 2012. Information and Communication Technologies for Development - Guest Editors' Introduction. IEEE Pervasive Computing, 11(3).

[CCK+05] Chen, W., Clarke, L., Kurose, J., and Towsley, D. Optimizing cost-sensitive trust-negotiation protocols. IEEE Computer and Communications Societies (INFOCOM), 2005.

[Cha12] S. Charney. Trustworthy Computing Next. Microsoft, white paper, 2012.

[CLM+02] Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., and Reagle, J. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. W3C Recommendation, 2002.

[CoC13] P. Collin, N. Colin. Mission d'expertise sur la fiscalité de l'économie numérique. Ministère des Finances et de l'Economie. Rapport au Ministre de l'économie et des finances, au Ministre du redressement productif, au Ministre délégué chargé du budget et à la Ministre déléguée chargée des petites et moyennes entreprises, de l'innovation et de l'économie numérique. Jan. 2013.

[Coo71] Cook, S. A. The complexity of theorem-proving procedures. In Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (1971).

[DDJ+03] E. Damiani, S. De Capitani Vimercati, S. Jajodia, S. Paraboschi, P. Samarati, 'Balancing Confidentiality and Efficiency in Untrusted Relational DBMSs', *ACM Conference on Computer and Communications Security (CCS)*, 2003.

[DDP+02] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, P. Samarati, 'A Fine-Grained Access Control System for XML Documents', *ACM Transactions on Information and System Security (ACM TISSEC)*, (5)2, 2002.

[DeS11] Debnath B., Sengupta S., Li J., "SkimpyStash: RAM Space Skimpy Key-Value Store on Flash," SIGMOD, 2011.

[DGM+07] Diao, Y., Ganesan, D., Mathur, G., and Shenoy, P. J. Rethinking data management for storage-centric sensor networks. In CIDR, pp. 22–31, 2007.

[Dir95] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data. Official Journal of the EC, 23, 1995.

[EhJ07] J. H. Ehlers and S. A. Jassim. Wavelet library for constrained devices. volume 6579, pages 65790P–65790P–11, 2007.

[Eur85] European Directive 95/46/EC, 'Protection of individuals with regard the processing of personal data', Official Journal L 281, 1985.

[Euro08] Eurosmart. Smart USB token. White paper, Eurosmart, 2008.

[FGK02] Fourer, R., Gay, D. M., and Kernighan, B. W. AMPL : A Modeling Language for Mathematical Programming, second edition. Duxbury Press, 2002.

[GiD14] Giesecke & Devrient. StarSign® Mobile Security Card SE 1.2, Datasheet, 2014.

[GoB14] J. González and P. Bonnet. Towards an open framework leveraging a trusted execution environment. In Cyberspace Safety and Security. Springer, 2013

[HAF+06] H.J.W. van Heerde, N. Anciaux, L. Feng, P. Apers. Balancing Smartness and Privacy for the Ambient Intelligence. First European Conference on Smart Sensing and Context (EuroSSC), Lecture Notes in Computer Science 4272 springer 2006, Enschede, The Netherlands, Oct. 2006

[HCL+97] Haas L. M., Carey M. J., Livny M., Shukla A., "Seeking the truth about ad hoc join costs," VLDB Journal, 1997.

[HFA09] H. van Heerde, M. Fokkinga, N. Anciaux. A Framework to Balance Privacy and Data Usability Using Data Degradation. IEEE International Conference on Computational Science and Engineering (CSE), Los Alamitos, CA, USA, 2009.

[IBM03] IBM corporation, 'IBM Data Encryption for IMS and DB2 Databases v. 1.1', 2003. http://www-306.ibm.com/software/data/db2imstools/html/ibmdataencryp.html.

[ITU11] ITU. 2011. The Role of ICT in Advancing Growth in Least Developed Countries – Trends, Challenges and Opportunities. http://www.itu.int/pub/D-LDC-ICTLDC.2011

[Kar72] Karp, R. M. Reducibility among combinatorial problems. In Complexity of Computer Computations (1972), pp . 85-103.

[Kha11] F. Khatibloo. Personal Identity Management - Preparing For A World Of Consumer-Managed Data. Forrester Report, Sept. 30, 2011.

[KoV11] Koltsidas I., Viglas S. D., "Data management over flash memory," SIGMOD, 2011.

[LFA11] Lim H., Fan B., Andersen D., Kaminsky M., "SILT: a memory -efficient, high-performance key-value store", SOSP, 2011.

[LiR99] Li, Z., and Ross, K. A., "Fast joins using join indices", VLDB Journal, 1999.

[LLL10] Li, F., Luo, B., & Liu, P. (2010, October). Secure information aggregation for smart grids using homomorphic encryption. In Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on (pp. 327-332). IEEE.

[Mer90] R. Merkle, 'A Certified Digital Signature', *Advances in Cryptology (Crypto'89)*, LNCS, vol.435, Springer--Verlag, 1990.

[MHV+08] 'Les Yvelines, acteurs et partenaires de la Recherche et Développement' lors de la Convention d'Affaires 'Les RDV Carnot, Palais des Congrès, Versailles. Table ronde animée par : Christian Beley, Sous-Directeur Pôle économique CG78, Frédérick Becquet, Chargé de mission R&D CG78. Participants : Pr. Luc Montagnier, PDG Nanectis Biotechnologies, Yan Haentjens, PDG Vectrawave, Jean-Pierre Arragon, Directeur Portfolio Management Continental Automotive, et Nicolas Anciaux, Chargé de recherche à Inria Paris-Rocquencourt. Mars 2008.

[Mil14] Claire Cain Miller. Revelations of N.S.A. Spying Cost U.S. Tech Companies. The New York Times, 21 Mars 2014.

[MLC+13] M. Madden, A. Lenhart, S. Cortesi, U. Gasser. Teens and Mobile Apps Privacy. Pew Internet and American Life Project. Août 2013.

[MOP+00] Muth P., O'Neil P., Pick A., Weikum G., "The LHAM log-structured history data access method", VLDB Journal, 2000.

[Mos05] Moses, T. Extensible access control markup language (xacml) version 2.0. Oasis Standard, 2005.

[MSW+14] Y.A. de Montjoye, E. Shmueli, S.S. Wang, A.A. Pentland. openPDS: Protecting the Privacy of Metadata through SafeAnswers. PloS one, 9(7), 2014.

[MWB11] A. D. K. Mulligan, L. Wang, and A. J. Burstein, "Final Project Report Privacy in the Smart Grid: An Information Flow Analysis," CIEE Report, 2011.

[NTB+12] A. Narayanan, V. Toubiana, S. Barocas, H. Nissenbaum, D. Boneh, D. A critical look at decentralized personal data architectures. arXiv preprint arXiv:1202.4503, 2012.

[OCG+96] O'Neil P., Cheng E., Gawlick D., O'Neil E., "The log-structured merge-tree (LSM-tree)", Acta Informatica, 1996.

[PBV+01] P. Pucheral, L. Bouganim, P. Valduriez, C. Bobineau, 'PicoDBMS: Scaling down Database Techniques for the Smartcard', *Very Large Data Bases Journal, VLDBJ, 10(2-3)*, 2001. Special issue on the best papers from VLDB'2000.

[Pri74] The Privacy Act, 5 U.S.C.§552a, 1974. http://www.usdoj.gov/04foia/privstat.htm

[Res14] Résolution législative du Parlement européen du 12 mars 2014 sur la proposition de règlement du Parlement européen et du Conseil relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (règlement général sur la protection des données). 12 mars 2014.

[RiD11] Rial, A., & Danezis, G. (2011, October). Privacy-preserving smart metering. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society (pp. 49-60). ACM.

[RoO92] Rosenblum M., Ousterhout J., "The Design and Implementation of a Log-Structured File System", ACM TOCS, 1992.

[SAB+07] C. Salperwyck, N. Anciaux, M. Benzine, L. Bouganim, P. Pucheral, D. Shasha. GhostDB: Hiding Data from Prying Eyes. 33th International Conference on Very Large Data Bases (VLDB), demo. paper, Vienna, Austria, Sept. 2007.

[Sam01] Samarati, P. Protecting respondents' identities in microdata release. IEEE TKDE, 13(6), 2001.

[SeL76] Severance D., Lohman G., "Differential files: their application to the maintenance of large databases". ACM TODS, 1976

[SmR] Schmid P., Roos A., "SDXC/SDHC Memory Cards, Rounded Up And Benchmarked", http://tinyurl.com/tom-sdxc

[Sta13] J. Staten. The Cost of PRISM Will Be Larger Than ITIF Projects. Forrester blog. Août 2013.

[Sun99] Sundaresan P., "General Key Index", US. Patent n° 5870747, 1999.

[TNP14] To, Q.-C., Nguyen, B., and Pucheral, P.: Privacy-Preserving Query Execution using a Decentralized Architecture and Tamper Resistant Hardware. EDBT 2014: 487-498.

[TsK07] Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 3(3), 1-13.

[TSW+08] C. C. Tan, B. Sheng, H. Wang, and Q. Li. Microsearch: When search engines meet small devices. In Pervasive Computing, volume 5013 of Lecture Notes in Computer Science, pages 93–110. Springer Berlin Heidelberg, 2008.

[TSW+10] C. Tan, B. Sheng, H. Wang, and Q. Li. Microsearch: A search engine for embedded devices used in pervasive computing. ACM Trans. Embed. Comput. Syst., 9(4):43:1–43:29, Apr. 2010.

[VaA09] Demonstration of Electronic Health Records (EHR) on Java Card 3.0 Based Devices. Jean-Jacques Vandewalle, Research Engineer GEMALTO, Nicolas Anciaux, Researcher (Inria). Smart Event 10th Edition, World e-ID 2009, Sophia-Antipolis, sept. 2009.

[Val87] P. Valduriez, 'Join Indices', *ACM Transactions on Database Systems (ACM TODS)*, 12(2), 1987.

[Vin02] R. Vingralek, 'Gnatdb: A small-footprint, secure database system', *28th International Conference on Very Large Data Bases (VLDB)*, August 2002.

[VMS02] R. Vingralek, U. Maheshwari, W. Shapiro, 'TDB: A Database System for Digital Rights Management', *8th International Conference on Extending Database Technology (EDBT)*, March 2002.

[VWA+12] Vo, H. T., Wang, S., Agrawal, D., Chen, G., & Ooi, B. C. (2012). LogBase: a scalable log-structured database system in the cloud. Proceedings of the VLDB Endowment, 5(10), pp. 1004-1015, 2012.

[War10] Fading data could improve privacy. By M. Ward, BBC News. 16 June 2010. http://www.bbc.co.uk/news/10324209

[WCK03] Wu C., Chang L., Kuo T., "An Efficient B-Tree Layer for Flash-Memory Storage Systems," RTCSA, 2003.

[WEF11] The World Economic Forum. Personal Data: The Emergence of a New Asset Class. Nov. 2011.

[WEF12] The World Economic Forum. Rethinking Personal Data: Strengthening Trust. May 2012.

[Wei02] Weininger, A., "Efficient execution of joins in a star schema", SIGMOD, 2002.

[WTL10] H. Wang, C. C. Tan, and Q. Li. Snoogle: A search engine for pervasive environments. Parallel and Distributed Systems, IEEE Transactions on, 21(8):1188–1202, Aug 2010.

[XiT06] Xiao, X., and Tao, Y. Personalized privacy preservation. In Proceedings of ACM SIGMOD, 2006.

[YFA+08] Yao, D., Frikken, K.B., Atallah, M.J., and Tamassia, R. Private information: To reveal or not to reveal. In ACM TISSEC, 12(1), 2008

[YGM08] Yan, T., Ganesan, D., and Manmatha, R. Distributed image search in camera sensor networks. In Proc. of the 6th ACM Conference on Embedded Network Sensor Systems, SenSys'08, pp. 155–168, 2008.

[YPM09] Yin S., Pucheral P., Meng X., "A Sequential Indexing Scheme for Flash-based embedded systems," EDBT, 2009.

[YSM+08] Yap, K.-K., Srinivasan, V., and Motani, M. Max: Wide area human-centric search of the physical world. ACM Transactions on Sensor Networks, 4(4):26:1-34, 2008.

[ZoM06] Zobel, J. and Moffat, A. Inverted files for text search engines. ACM Computing Survey, 38(2), 2006.

# Annexe A.

# Secure Personal Data Servers: a Vision Paper

Tristan Allard, Nicolas Anciaux, Luc Bouganim, Yanli Guo, Philippe Pucheral, Benjamin Nguyen, Lionel Le Folgoc, Indrajit Ray, Indrakshi Ray, Shaoyi Yin

*Proceedings of the VLDB Endowment (PVLDB), Volume 3(1), pp. 25-35, 2010.*

# Annexe B.

# Trusted Cells: a Sea Change for Personal Data Services

Nicolas Anciaux, Philippe Bonnet, Luc Bouganim, Benjamin Nguyen, Philippe Pucheral, Iulian S. Popa

*Conference on Innovative Database Reasearch (CIDR), 4 pages, 2013.*

# Annexe C.

# Folk-IS: Opportunistic Data Services in Least Developed Countries

Nicolas Anciaux, Luc Bouganim, Thierry Delot, Sergio Ilarri, Leïla Kloul,
Nathalie Mitton, Philippe Pucheral

*Proceedings of the VLDB Endowment (PVLDB), Volume 7(5), pp. 425-428, 2014.*

# Annexe D.

# MILo-DB: a personal, secure and portable database machine

Nicolas Anciaux, Luc Bouganim, Philippe Pucheral, Yanli Guo, Lionel Le Folgoc, Shaoyi Yin

*Distributed and Parallel Databases (DAPD), Volume 32(1), pp. 37-63, 2014.*

# Annexe E.

# Limiting Data Collection in Application Forms

## A real-case Application of a Founding Privacy Principle

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis

*International Conference on Privacy, Security and Trust (PST), pp. 59-66, 2012*

# Annexe F.

# Curriculum Vitae – Nicolas Anciaux