Identité numérique et anonymat : Concepts et mise en oeuvre

El Hassan Bezzazi
IREENAT, Université de Lille 2

Introduction

Le travail interdisciplinaire mené dans le programme de recherche Asphalès a été l'occasion pour des chercheurs issus du monde du droit et de celui de l'informatique de commenter un certain nombre de concepts et d'en débattre. Parmi ces concepts, liés à la sécurité juridique et à la sécurité informatique, le concept de donnée personnelle a été le concept qui nous a le plus percuté. Sans doute parce que nous nous sentons immédiatement concernés mais aussi à cause du fait qu'il soit transversal à d'autres concepts considérés tels que l'intégrité, l'archivage ou l'effacement. Le concept de donnée personnelle est au coeur de la notion d'identité numérique et de son dual qu'est l'anonymat. La définition de l'identité numérique reposera donc sur la définition d'une donnée personnelle pour laquelle il existe une référence juridique dans la loi. L'anonymat est important, parfois nécessaire, pour la protection de la vie privée. Son concept est intimement lié à celui de l'identité. En effet, l'anonymat est souvent défini comme étant l'état d'être non identifiable dans un ensemble de sujets. Ces concepts prennent de nouvelles dimensions dans le monde numérique où il est question de la personnalité numérique par opposition aux personnalités publique et privée de l'individu. Globalement et sommairement, l'identité numérique d'une personne physique serait la somme de ses données personnelles dans le monde numérique. Il devient alors nécessaire de recenser les différents intérêts suscités par ces notions ainsi que leur mise en oeuvre pour approcher une ontologie de l'identité numérique qui profitera, entre autres, à leur définition juridique. Nous nous proposons dans cet article de procéder à une approche descendante et modulaire dans la définition d'un modèle pour appréhender ces deux notions et certaines autres qui viennent se greffer autour d'elles. L'efficacité du modèle dépend de sa capacité à intégrer les éléments de la réalité qui sont pertinents pour son utilisation. Comme tout modèle, notre modèle sera réducteur à commencer par la vision d'un individu comme un ensemble de données. L'objectif de cet article est d'avancer une pseudo formalisation du concept de l'identité numérique qui pourrait ouvrir la voie à la mise en place d'une ontologie formelle de ce concept partageable par les différents acteurs autour de ce concept. Un certain nombre de propositions ont été faites dans la littérature [Cam06, Cla94, HWV03], notre proposition s'en distingue tant au niveau du formalisme utilisé qu'au niveau de l'appréhension globale des concepts et des relations connexes.

L'identité numérique dans son approche juridique

D'après la loi "Informatique et libertés":
« Constitue une donnée à caractère personnel toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro

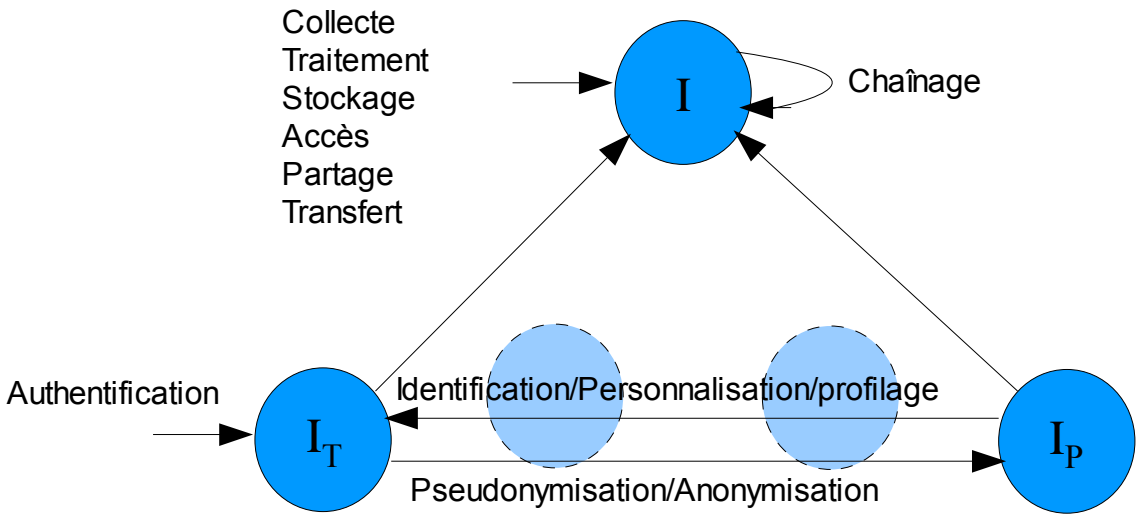
d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne. » (art. 2)
« La personne concernée par un traitement de données à caractère personnel est celle à laquelle se rapportent les données qui font l'objet du traitement. » (art. 2)
La définition distingue deux cas:

- le cas d'une personne identifiée et dans ce cas toute information le concernant qui vient s'ajouter à son identifiant (simple ou composé) est considérée comme une donnée à caractère personnel.
- le cas d'une personne X qui n'est pas identifiée et pour laquelle on dispose d'un ensemble de données parmi lesquelles certaines données suffisent à l'identifier éventuellement à l'aide de moyens dont peut disposer une personne cherchant à l'identifier.

La notion d'identifiant, le nom par exemple, n'est pas liée à sa nature seulement mais aussi au contexte implicite. En effet, si le médecin d'une commune de moins de 500 habitants est plus facilement identifiable qu'un médecin dans une commune de plus d'un million d'habitants. Ainsi le nom ne constitue pas dans l'absolu nécessairement un identifiant mais le contexte ou sa valeur peuvent en faire un. On peut alors considérer le contexte et la valeur comme des moyens implicites permettant ou non l'identification d'une personne le long d'un certain processus d'inférence. A l'inverse, une donnée est anonyme quand l'établissement d'un rapport à une personne physique exige la mise en oeuvre d'un ensemble de moyens techniques, financiers, humains de manière déraisonnable. L'identité numérique, si elle doit être unique, peut être conçue comme l'ensemble de toutes ses identités numériques partielles. Chacune de ces identités partielles est supposée être cohérente. Sans moyen de chaînage de deux identités partielles, un observateur – qui peut être un programme- distinguera deux individus. Si l'observateur arrive à chaîner les deux identités, il devra en tirer une synthèse cohérente car des faits peuvent se contredire. Une approche prudente exclura tout simplement de tels faits. Ces identités numériques peuvent être crées par une personne, un organisme ou un programme et peuvent ne faire référence à aucune personne physique réelle quand bien même un lien entre elle et son créateur est établi. Cependant, dans ce dernier cas cette identité pourra être considérée comme une donnée de son auteur au même titre qu'un billet dans un blog décrivant son auteur. Ceci étant, nous élaboreront notre étude dans cet article sur la base du cercle restreint des identités partielles dont les faits sont vrais, éventuellement après nettoyage, pour une personne physique réelle.

Un modèle pour les identités numériques

Notre modèle est, selon une démarche descendante, d'abord une ontologie pseudo formelle et sera complété dans le paragraphe suivant par une définition logico-ensembliste de l'identité numérique. La figure ci-dessous donne une vue globale du modèle de l'identité numérique pour une personne donnée X. La classe I représente l'ensemble de ces identités qui pourront être classées à nouveau dans l'une des deux sous classes disjointes, la classe I_T des identités totales qui permettent l'identification de X et la classe I_P des identités partielles qui ne permettent pas cette identification. La figure met aussi en évidence les principales opérations qui s'appliquent sur les identités numériques.



Notons qu'au niveau de l'implantation, les identités I peuvent être structurés ou non. Elles sont structurées quand elles sont issues de bases de données tels que les fichiers logs. Elles le sont moins quand elles sont issues d'un texte tel que le contenu d'une page web. La représentation logico-ensembliste que nous proposerons pour une identité numérique fera abstraction de ces détails.

Le **dépôt** des données pour alimenter l'identité peuvent provenir de la personne physique ou de sources externes. Même quand elles proviennent de la personne, celle-ci peut ne pas en être consciente, c'est typiquement le cas des traces. Le **chaînage** de deux identités peut être rendu possible en raison d'informations communes qui en sont issues et qui sont réputées uniques telle que l'adresse email, l'adresse ip ou un cookie. L'opération d'identification fait intervenir un ensemble de connaissances, celles de l'observateur, et comme nous le verrons peut aboutir ou non en ne permettant que la personnalisation ou le profilage des individus. La **collecte, le stockage , le partage, le transfert** des données à caractère personnel et leur **traitement** sont encadrés par la loi, notamment par rapport au consentement, à la finalité, à la proportionnalité et à durée de rétention. En particulier, la finalité doit être licite, déterminée, explicite et légitime. Le caractère légitime est déterminé par le consentement de la personne concernée ou une nécessité démontrée. Ces actions échapperont à la réglementation si les données , par exemple à la suite d'une anonymisation, ne permettent pas l'identification par tout moyen raisonnable, par chaînage ou inférence par exemple. L'**accès** aux données peut être protégé ou non. Dans le cas où il est protégé, les personnes autorisées doivent s'identifier et procéder à une **authentification**. Ils peuvent avoir des droits différents.

Une définition logico-ensembliste

I(x) est l'ensemble des assertions vraies pour l'individu x et Id(x) le sous-ensemble de I(x) constitué des faits qui l'identifient. Soit D(x) un sous ensemble de I(x)-Id(x), il est dit identifiant

de x pour un observateur a si les connaissances générales - et non spécifiques à x dans la mesure où ces données sont dans D(x)-K(a) - de a lui permettent d'identifier x:
$K(a) \cup D(x) \rightarrow i(x)$ avec $i(x) \in Id(x)$. Le symbole $\rightarrow$ désigne une inférence générale pouvant prendre plusieurs formes: déductive, inductive, abductive ou probabiliste. Il doit être clair que dans ce cas $K(a) \cup D(x) \rightarrow i(y)$ avec $i(y) \in Id(y)$ pour $x \neq y$. Remarquons cependant que cette identification n'est pas infaillible dans la mesure où les connaissances de a peuvent ne pas être assez suffisantes pour lui permettre de noter que d'autres individus vérifient D(x).
Si D(x) n'est pas identifiant de x pour a, deux cas se présentent:

- Les connaissances K(a) de a ne sont pas suffisantes ou il ne dispose pas de moyens suffisants pour identifier x
- Ses connaissances lui permettent de recenser plusieurs individus pour lesquels D(x) est vrai

Le processus d'identification peut se présenter sous deux formes:

- à partir de D(x) pour trouver x. Par exemple remonter à une personne à partir de fichiers logs.
- à partir d'un x connu en faisant varier D(x). Par exemple retrouver une ancienne connaissance sur le web en utilisant un moteur de recherche.

Les processus que nous venons de voir sont subjectifs. Leur objectivation consiste à idéaliser tous les observateurs dont le processus d'identification est infaillible et disposent tous des mêmes moyens nécessaires pour identifier x sur la base de D(x) si celui-ci suffit à identifier x dans l'absolu, c.à.d. s'il n'existe pas deux individus distincts pour lesquels D(x) est vrai. Typiquement, ceci est représenté par l'utilisation d'une base de connaissances K commune à l'ensemble des observateurs. Par ailleurs; il est intéressant de noter que, à notre connaissance, la définition de l'anonymat rencontrée dans la littérature est basée sur le premier processus. Pourtant une personne que je connais et dont j'ai perdu la trace depuis quelques années devient anonyme pour moi. Identifier cette personne représentera une tâche plus ou moins délicate. A supposer que cette personne ait une identité numérique sur le web, il s'agira pour moi de fournir au moteur de recherche des éléments d'information issus de mes souvenirs sur cette personne pour espérer par discriminations successives (nom approximatif, prénom, lycée, ville, compétences...) et feedback (photo) l'identifier. Une mesure pour l'anonymat dans ce cas est à proposer. On peut même envisager un système expert qui aide sur la base d'une telle mesure l'utilisateur à isoler une identité numérique éventuelle pour une personne perdue de vue. L'intérêt qui est porté plus sur le premier processus peut s'expliquer par sa pertinence pour les entreprises qui procèdent à des profilages ou à des personnalisations.

Quantifier l'anonymat

L'utilisateur peut avoir plusieurs identités numériques mais certaines données personnelles divulguées dans ces profils telles que l'adresse email peuvent servir à coupler les profils d'un même utilisateur ouvrant ainsi la voie à l'identification. Dés lors l'utilisateur doit être à même de mesurer la capacité d'un service à assurer la protection de ces données personnelles et de son anonymat. Une telle quantification est relativement possible à l'aide, par exemple, de la théorie de l'information comme le propose [Dia02] dans le cadre d'un système à N utilisateurs. Chaque utilisateur u_j génère un ensemble de requêtes. Soit R le nombre total des requêtes produites sur un intervalle de temps donné et $(R_1,...,R_N)$ l'ensemble des requêtes. L'analyse de la base de données permet à l'attaquant d'affecter à chaque utilisateur une probabilité sur son lien avec une requête donnée R_j. Le degré maximal d'anonymat relativement à cette requête a lieu quand un attaquant

juge équiprobable la pertinence de cette requête pour tous les utilisateurs. Le concept d'entropie permet de mesurer l'information contenue dans cette distribution de probabilités et sera donc utilisée pour calculer le degré d'anonymat présenté par les utilisateurs. L'entropie du système après analyse $H=-\Sigma_{(1,N)}\ p_i\log(p_i)$ est comparée à l'entropie maximale $H_M=\log_2(N)$ pour calculer la quantité d'information gagnée par l'attaquant H_M-H qui lui permet une certaine distinction entre les utilisateurs par rapport a la requête R_j. Le degré d'anonymat pour la requête R_j est alors $dj=1-(H_M-H)/H_M=H/H_M$. Le calcul de la moyenne de ces degrés définit le degré d'anonymat de l'ensemble du système analysé.

Les acteurs

Les concepts d'identité et d'anonymat numériques ainsi définis, nous allons maintenant aborder le contrôle et la marge de manoeuvre ainsi que leurs déterminants que peut avoir chacun des acteurs en rapport avec ces concepts. Nous avons distingué principalement cinq groupes d'acteurs autour du concept de l'identité numérique. Ils n'utilisent pas toujours les mêmes termes et n'ont pas toujours la même acception d'un même terme, d'où l'intérêt de la définition d'une ontologie médiatrice. Cette formalisation devra pour chaque groupe clarifier ses concepts et les rapprocher à d'autres concepts dans les autres domaines (identité, anonymat, données personnelles, finalité, proportionnalité…). La typologie des acteurs que nous proposons est sans doute perfectible. Par exemple nous aurions pu distinguer la classe des acteurs «attaquant» qui agissent à l'encontre de la confidentialité de l'identité numérique et de l'anonymat. Nous pensons qu'une telle classe n'est pas réellement opportune dans la mesure où ces attaques peuvent être potentiellement le fait de n'importe quelle autre classe d'acteurs de notre typologie.

Les utilisateurs

L'ontologie doit éclairer l'utilisateur et l'assister dans ses différents rôles dans le monde numérique pour une meilleure appréhension de son identité numérique à défaut d'un réel contrôle. Les métaphores sont souvent utilisées à cet effet, comme par exemple le système à tiroirs prévu pour l'e-administration en France ou le portefeuille des cartes d'identité dans Cardspace [Cha06]. Les données à caractère personnel peuvent être collectées de façon explicite ou implicite. L'individu exhibe un comportement sur le net à commencer par le pseudo utilisé, la réactivité, la fréquence et la nature de ses contributions et leur style. Le cumul de ces signaux aboutit dans le temps à une approximation de la personnalité. Dans un contexte commercial, le profilage est la première fonction consommatrice de ces données et n'a à priori pas besoin de l'identité mais seulement de la personnalité. Notons alors le fait qu'il n'y ait pas d'identification directe ou indirecte annule le caractère personnel des données collectées au sens de la réglementation informatique et libertés. L'importance de l'anonymat pour l'utilisateur est à relativiser en fonction de sa culture, sa nature, son niveau intellectuel, ses compétences et de son activité numérique qui peut être licite ou illicite. L'utilisateur averti pourra recourir à des outils d'anonymisation dans sa navigation et prendra un certain nombre de précautions lors de son inscription à un service en ligne en utilisant des pseudonymes ou des adresses email jetables et en étant vigilant quant aux faux sites pratiquant le phishing. L'utilisateur doit être conscient que la possession et l'utilisation d'outils d'anonymisation dont légales, c'est l'activité entreprise qui peut être illégale. Enfin l'utilisateur doit s'informer sur ses droits en rapport avec sa vie privée en se référant notamment à des services dont le but premier est de l'assister en la matière tels que ceux offerts par la CNIL.

Les organismes

Qu'ils soient des entreprises, des administrations ou autres associations, les organismes doivent proposer des ontologies claires sur leur politique et leur engagement relatifs à la collecte et au traitement des données personnelles à défaut de se situer par rapport à une ontologie partagée. Ils doivent informer l'utilisateur sur l'existence d'un droit d'opposition à la diffusion de ses renseignements, d'un droit d'accès, de modification et de radiation de ces renseignements et la manière de les exercer. Le fournisseur de service doit aussi porter à la connaissance de l'utilisateur le niveau de sécurité du site web et afficher la ou les législations auxquelles il est assujetti. Enfin, le recours éventuel à un sceau de certification ou à une technologie de protection de la confidentialité [pri] accroîtra la confiance dans le site. L'importance du respect de la confidentialité pour une entreprise ou un gouvernement est à relativiser en fonction des lois nationales à laquelle elle est soumise, des moyens dont elle dispose, techniques (logiciels et protections adaptées) ou humains (correspondants CNIL). Les pouvoirs publics doivent à la fois veiller au respect du droit et à la dynamique de l'économie.
Par ailleurs les données personnelles collectées par ces services peuvent être utilisées par des tiers dans des buts louables telles que des statistiques pour optimiser la production des médicaments par exemple. L'anonymisation doit alors être mise en oeuvre ainsi que des opérations préservant la confidentialité. Cette anonymisation a cependant ces limites dans certains domaines tels que la publication des décisions de justice dans la mesure où les moteurs de recherche peuvent servir à l'identification éventuelle des partis. En effet les faits divers de la presse locale relatent souvent les événements liés aux décisions de justices et permettent facilement, au moyen des moteurs de recherche, de lever l'anonymat.

Les juristes

Le problème de la clarté et de l'accessibilité des textes juridiques est essentiel pour la sécurité juridique et les concepts techniques utilisés ne semblent pas toujours bien mesurés. Les interprétations qu'en peuvent donner les juges, l'entreprise, le simple citoyen mais aussi le gouvernement dans la rédaction des décrets d'application doivent coïncider autant que possible pour un meilleur respect de la norme. Pour ce faire, il est important qu'une expertise globale et pertinente précède l'adoption de nouvelles normes utilisant des termes techniques. Les atteintes à la vie privée peuvent venir de la législation même souvent justifiée pour des raisons de sécurité.
L'importance du respect de la confidentialité pour le législateur est fonction de ses motivations économique et sécuritaire et des contraintes posées par la hiérarchisation des normes. Dans le cas du stockage des logs et de la vidéosurveillance prévus dans la loi contre le terrorisme c'est l'opposition de deux droits et la façon de les concilier qui interpellent:

- le droit de l'individu à la protection contre le terrorisme et le devoir d'un état à le protéger
- le droit de l'individu à la protection des données personnelles et à la vie privée et le devoir d'un état à les protéger.

L'article 8 de la convention Européenne des droits de l'homme et des libertés fondamentales reconnaît à toute personne le droit au respect de sa vie privée et familiale, de son domicile et de ses correspondances. Cet article protège aussi l'individu contre des intrusions arbitraires des pouvoirs publics dans sa vie privée. Elles ne sont tolérées qu'en tant que mesure exceptionnelle explicitement prévue dans la loi et nécessaire dans une société démocratique sur des sujets touchant à l'ordre public et aux droits de l'individu.

Les développeurs d'applications

Le développement technologique devrait se faire en respectant le droit existant. Les normes technologiques qui sont souvent définies à posteriori doivent alors formaliser la conformité avec le droit.
Sur la base des ontologies ci-dessus, les développeurs seront à même d'offrir aux entreprises un produit de qualité au niveau de la légalité de son exploitation et de l'ergonomie de l'interface homme-machine. Le W3C a développé une plateforme P3P (platform for privacy preferences) pour mettre en place un standard répondant aux préoccupations en matière de vie privée lors d'échange entre internautes et sites web ou encore entre sites web dans le cas des web services. Le but est que l'utilisateur ait plus d'information et de contrôle sur les données personnelles collectées par les sites visités donc une meilleure transparence sur la base d'un langage standardisé définissant les questions relevant de la vie privée. La politique en matière de vie privée du site ayant adhéré au standard P3P est codée dans un fichier XML lisible par les navigateurs intégrant ce standard. Ce fichier fournit la réponse à un certain nombre de questions sur la collecte ou non des données personnelles. Le navigateur vérifie alors si ses réponses satisfont les préférences de l'utilisateur pour l'en informer et lui donner la possibilité éventuellement d'un réglage ad hoc pour pouvoir visiter le site mais en connaissance de cause. Sur le plan pratique, une organisation va mettre sur son site sa politique en matière de vie privée au format XML. Elle doit indiquer entre autres quelles sont les informations collectées, comment elles sont utilisées, la durée de conservation et qui a accès à l'information. Pour une prise en compte rationnelle de la protection de la confidentialité dans les applications, les développeurs de logiciels doivent se référer aux normes prévues à cet effet dans les critères communs et qui ne cessent d'évoluer [Tro02]. Par exemple, dans le développement d'un logiciel d'anonymisation pour une application donnée des choix devront être opérés selon une démarche méthodique respectant les normes pour assurer la réversibilité, l'irréversibilité ou l'inversibilité des informations [ADTC04].

Les défenseurs

Dans ce groupe, on mettra pèle-mêle les autorités indépendantes telles que la CNIL, des organismes de défense des consommateurs; les forums d'internautes, et surtout les développeurs de logiciel, notamment libre, dédié à la protection de la confidentialité et des personnes qui mettent à dispositions des ressources dédiées à cette même cause comme des serveurs d'anonymisation [tor]. Il convient de signaler dans ce paragraphe le projet Européen en cours EuroPriSe impliquant neuf partenaires Européens dont la CNIL pour mettre en place un sceau européen en faveur des produits des technologies de l'information qui auront fait preuve du respect de la confidentialité. Le produit obtiendra le label à l'issue d'une procédure de certification [eur].

L'authentification

Si l'anonymat est, à différents degrés, une revendication légitime pour un certain nombre de services, l'identification l'est tout aussi bien pour d'autres services qui l'exigent tels que l'e-administration et l'e-santé. Des solutions informatiques pour la gestion de l'identité numérique qui assurent son intégrité et sa confidentialité doivent alors être mises en oeuvre. Les solutions U-Prove SDK de Credentica, IDEMIX et dans une moindre mesure Cardspace en sont les meilleurs exemples [cre, ide, CH02, Cha06]. Par exemple le produit U-Prove permet aux organismes de

protéger les assertions liées à l'identité tout en offrant la possibilité à l'utilisateur, à l'aide de fonctions de contrôle et d'une authentification forte sans l'implication en temps réel d'un fournisseur d'identité, de prouver des assertions qui n'étaient pas anticipées et le transfert de données entre comptes non chaînables. Pour assurer une divulgation sélective des assertions ils mettent en oeuvre des algorithmes cryptographiques conséquents [Bra00, BCL04] pour implanter par exemple la signature aveugle et le protocole de la preuve à divulgation nulle de connaissances.

Conclusion

Nous avons proposé dans cet article quelques éléments pour la modélisation des concepts liés à l'identité numérique en rapport avec ses acteurs. Dans le cas de l'authentification, il convient de retenir qu'un système d'authentification qui protège la confidentialité est un système qui permet d'exprimer des assertions vérifiables, minimales et non chaînables. La définition d'une ontologie formelle sera opportune à la fois pour disposer d'un support permettant le développement de sa définition et pour situer et comparer les acteurs par rapport au respect de la confidentialité.

Références

[ADTC04] A. Abou El Kalam, Y. Deswarte, G. Trouessin, E. Cordonnier. Une démarche méthodologique pourl'anonymisation de données personnelles sensibles, Actes du 2èmeSymposium sur la Sécurité des Technologies de l'Information et des Communications (SSTIC 2004) 2004.

[BCL04] E. Bangerter, J. Camenisch, A. Lysyanskaya. A cryptographic framework for the controlled release of certified data. Twelfth International Workshop on Security Protocols, 2004.

[Bra00] S.A. Brands. Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy. MIT Press, 2000.

[Cam06] K. Cameron. The Laws of Identity. 2005.

[CH02] J. Camenisch, E. Van Herreweghen. Design and implementation of the idemix anonymous credential system, citeseer.ist.psu. edu/camenisch02design.html, 2002.

[Cha06] D. Chappell. http://msdn2.microsoft.com/en-us/library/aa480189.aspx, 2006.

[Cla94] R. Clarke. The Digital Persona and its Application to Data Surveillance, 1994 http://www.anu.edu.au/people/Roger.Clarke/DV/DigPersona.html

[cre] http://www.credentica.com/.

[Dia02] C. Diaz, S. Seys, J. Claessens, B. Preneel, Towards measuring anonymity, LNCS 2482, 2002

[eur] http://www.european-privacy-seal.eu/

[HWV03] G. Hogben, M. Wilikens, I. Vakalis. On the ontology of Digital Identification. Springer LNCS 2889, 2003.

[ide] http://www.zurich.ibm.com/security/idemix/

[Lau07] B. Laurie. Selective disclosure. http://www.links.org/files/selective-disclosure.pdf. 2007.

[pri] PRIME - Privacy and Identity Management for Europe, https://www.prime-project.eu/.

[tor] Tor: Un système de connexion anonyme à Internet. http://www.torproject.org/index.html.fr.

[Tro02] G. Trouessin. L'évolution des normes de sécurité. La lettre d'Adeli N° 49, 2002.