La Gazette des archives

Open data et archivage électronique, quelles convergences ?

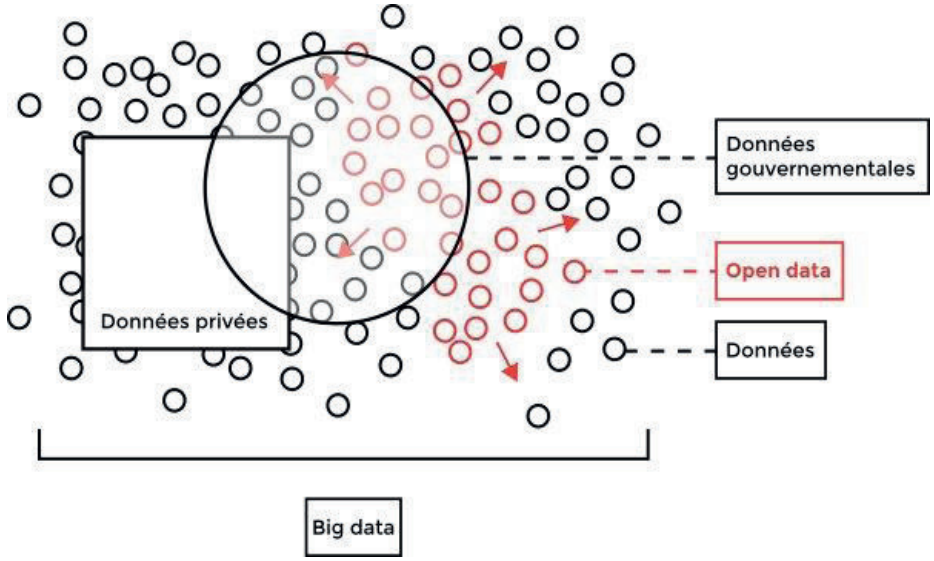
Céline Guyon, Cyril Longin, Jean-Daniel Zeller

Open data et archivage électronique, quelles convergences ?

Céline GUYON
Cyril LONGIN
Jean-Daniel ZELLER

Open data : concepts de base

Le concept d'*open data* (données ouvertes) a vu le jour au milieu des années 2000, prioritairement dans les pays anglo-saxons, en corrélation avec le concept d'*open governement*, promu à la suite de l'élection du président Obama en 2007. L'idée sous-jacente était celle d'un déficit démocratique qui pouvait être comblé par la communication plus proactive et plus fluide des données en possession de l'administration vers les citoyens/usagers. Un deuxième aspect est venu s'y ajouter qui est celui d'une plus-value que le traitement de ces données pouvait apporter aux entreprises qui les utiliseraient. Ce deuxième axe a quelque peu brouillé le sens attribué aux *open data* ce qui nécessite une clarification. Le schéma ci-après résume le champ traité par les *open data* et les autres termes utilisés actuellement dans ce domaine.



Relations entre les différents types de données © Sophie Czich (CC) BY-NC-SA

Les données privées, ou données personnelles, sont des données protégées par les législations sur la protection des données personnelles et n'entrent par principe pas dans le monde des données ouvertes.

Les données gouvernementales sont les données publiques potentiellement diffusables sous le nom d'*open data*.

Les autres données sont celles détenues par des entreprises ou des particuliers qui peuvent potentiellement également être offertes en tant qu'*open data*.

Les mégadonnées ou *big data* sont toutes les données volumineuses gérées de manière publique ou privée dans lesquelles figurent des *open data* mais aussi une multitude de micro-données collectées par toutes les applications informatiques principalement sur Internet.

Les *open data* sont les données publiques ou privés proposées aux utilisateurs selon certaines conditions.

Ces conditions ont été définies initialement lors d'une rencontre à Sebastopol (Californie), puis complétées dans l'article publié par la Sunlight Foundation « *Ten principles for opening up governement information* » c'est-à-dire « dix principes pour l'ouverture de l'information gouvernementale »[1]. Ces principes précisent que les données sont ouvertes quand elles répondent simultanément à ces dix critères (décrits plus loin).

[1] https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

Open data et transparence administrative

On peut voir dans les *open data* une continuité avec les principes de la transparence administrative concrétisés par les lois sur l'accès aux documents et à l'information promulguées dans la plupart des pays occidentaux ces dernières décennies[1].

Il existe cependant une différence de taille car la transparence suppose la communication de « documents » préconstitués communiqués à la demande du citoyen (modalité *pull*) alors que les *open data* consistent en la mise à disposition de jeux de données qui à la base ne sont pas des documents et doivent être rendus manipulables par l'utilisateur (modalité *push*). On passe donc d'une réponse à la demande à une publication proactive.

Par ailleurs, si les principes exposés ci-dessous sont légitimes, ils se heurtent dans les faits à des contraintes techniques non négligeables. Sans entrer dans les détails (chacun des principes exposés nécessiterait un débat en soi), on peut constater que les données accumulées par les services de l'administration ne l'ont pas été en vue de leur publication. Certaines l'ont été en vertu de prescriptions législatives ou réglementaires, d'autres pour des raisons de gestion pratique, d'autres parfois pour des raisons non identifiées. La grande majorité n'a pas fait l'objet d'analyse qualité ce qui est un des freins majeur à leur « ouverture », les administrations répugnant à diffuser des données dont la qualité et par conséquent « l'interprétabilité » ne peut être garantie. Curieusement, cette notion de fiabilité des données n'est pas inscrite dans les principes de Sébastopol, alors qu'elle constitue la condition *sine qua non* de leur crédibilité.

Open data et archives

Tel que cela est formulé ci-dessus, les *open data* semblent ne pas concerner les archives au sens historique mais plutôt les documents/données d'activité courants. Mais en y regardant de plus près, chacun des principes interroge la pratique archivistique.

- Les données doivent être complètes

La notion de complétude ne peut être générale. Elle doit se rapporter à un domaine d'activité précis : les données doivent être complètes par rapport à….

[1] On trouvera une liste de référence à ces lois dans l'article wikipedia anglais suivant : https://en.wikipedia.org/wiki/Freedom_of_information_laws_by_country

ou alors se rapporter au fait que l'on n'a pas retranché une partie du jeu de données initiales. Dans les deux cas, les notions archivistiques d'intégrité des fonds et de respect de l'ordre initial trouvent leur application ici. À la nuance près qu'une base de données n'est que très rarement close.

▪ Les données doivent être primaires (c'est-à-dire brutes)

Ce principe est une vue naïve qui présuppose que la donnée est « donnée » (c'est-à-dire le reflet d'une réalité objective abstraite. Or, comme l'explique Bruno Latour[1], la donnée est « obtenue » c'est-à-dire qu'elle est l'objet dès son origine d'un processus de fabrication qui n'est connaissable que par l'ajout de métadonnées qui documente justement sa source. En archivistique, on appelle cela le contexte.

▪ Les données doivent être fraîches

Entendez : elles doivent être à jour, ou mises à jour. Ici se trouve une des contrainte des *open data* qui doivent être pertinentes dans le temps. Alors que les archives sont pertinentes « hors du temps » ou plutôt indépendamment de la chronologie. Idéalement, cela nécessiterait que chaque donnée soit « datée » pour être exploitable à long terme (cela est vrai dans la cadre d'application de type « entrepôt de données », mais la plupart des applications dont sont issues les *open data* ne répondent pas à ce cas de figure).

▪ Les données doivent être accessibles

C'est est un truisme dans l'hypothèse de l'*open data*.

▪ Les données doivent être électroniquement lisibles par une machine

C'est la condition *sine qua non* de leur ré-exploitabilité et un obstacle majeur à la publication de données sous la forme de tableau en tant que document.

▪ Les données doivent être accessibles sans discrimination

On rejoint ici l'exigence de la transparence, qui s'applique initialement aux documents. Dans le cadre des données, il existe cependant une autre limite à l'accessibilité : la capacité (technique et intellectuelle) de les traiter.

▪ Les données doivent être disponibles sous des formats ouverts (dont les spécifications techniques sont publiques et sans restriction d'accès)

C'est la condition complémentaire au principe précédent.

▪ Les données doivent être disponibles sous une licence libre

Ceci est une condition complémentaire du principe précédent.

---

[1] *Face à Gaïa: Huit conférences sur le nouveau régime climatique*, Bruno Latour, 2015, 2e conférence, note 14.

▪ Les données doivent être accessibles de façon pérenne en ligne

Cette condition implique que l'archivage électronique des données soit assumé. Ce qui est encore loin d'être le cas des données actuellement proposées. On se retrouve dans deux cas de figure :

- soit les données ont une faible durée de vie (horaire de bus par exemple) et n'ont pas pour vocation à être conservées dans la durée ;

- soit les données sont dignes d'être archivées (séries statistiques par exemple) et les outils mis en place par les Archives devraient permettre également leur publication en mode *open data* (voir plus loin).

▪ Les données doivent être sans coût d'utilisation (le prix de mise à disposition ne doit pas excéder le coût de production)

Ce principe ouvre le débat sur la nature des licences attachées aux *open data* et sur la nature du coût attaché à la constitution des données publiques[1].

Open data et processus d'archivage électronique

De prime abord et dans une approche un peu caricaturale, tout oppose *open data* et processus d'archivage : d'un côté, l'immédiateté de l'*open data* et de l'autre, le temps infini de l'archivage. À y regarder de plus près, l'*open data* interroge la place de l'archivage dans le nouvel écosystème numérique.

On parle d'*open data* lorsque les données ont été publiées sur des plateformes dédiées car elles deviennent accessibles à tous. Le processus de publication des données publiques fait largement écho au processus d'archivage.

Dans ses principes d'abord. L'*open data* s'inscrit dans une tendance qui considère l'information publique comme un bien commun dont la diffusion est d'intérêt public et général. Les archives appartiennent au domaine public mobilier et leur conservation « est organisée dans l'intérêt public tant pour les besoins de la gestion et de la justification des droits des personnes physiques ou morales, publiques ou privées, que pour la documentation historique de la recherche »[2]. Les valeurs partagées par les partisans de l'*open data* et les archivistes sont similaires, les finalités de l'ouverture des données publiques et de l'archivage sont proches : transparence administrative, libre accès aux sources, respect de la vie privée et propriété collective des données.

[1] Voir le rapport Trojette en bibliographie.
[2] Art L. 211-2 du Code du patrimoine.

Dans ses modes opératoires ensuite. La publication des données publiques suppose d'une part la description des données à publier à l'aide d'un jeu de métadonnées puis le transfert des données et de leurs métadonnées vers une plateforme dédiée. L'expression « versement » est d'ailleurs utilisée dans le *vademecum* sur l'ouverture et le partage des données publiques[1] pour qualifier l'opération de publication de jeux de données sur la plateforme data.gouv.fr. *A contrario*, il est intéressant de noter que l'expression « transfert » a tendance à supplanter l'expression « versement » dans le nouveau vocabulaire de l'archivage électronique. Les dispositifs de versement des données sur la plateforme française data.gouv.fr et de versement des archives dans un système d'archivage électronique sont similaires avec d'un côté, le versement manuel (adapté aux versements ponctuels) et de l'autre, le versement automatisé (adapté aux versements réguliers).

Dans ses moyens également. Les données, pour être publiables, doivent être dans des formats ouverts et non propriétaires et les jeux de données décrits à l'aide de métadonnées. L'ajout de métadonnées permet de contextualiser les données (qui a produit les données ? Quand les données ont-été produites ? Quelle est la période temporelle concernée ? Quelles sont les zones géographiques couvertes ? Quelles sont les thématiques des données ?), et d'en préciser le contenu. Ces opérations de qualification et d'indexation des données sont décrites comme « une étape essentielle pour faciliter la réutilisation des données publiques »[2]. Le *vademecum* sur l'ouverture et le partage des données publiques insiste sur la nécessité des opérations de qualification des données et métadonnées et encourage les producteurs à « intégrer la perspective de l'ouverture des données et le besoin de qualification des jeux de données dans la conception et la rénovation des systèmes d'information »[3]. Les données publiées sont des données « brutes » mais leur valeur ajoutée, du point de vue de l'objectif de réutilisation, naît de l'ajout de métadonnées pertinentes et normalisées. De la même manière, pour être archivés, les données ou documents doivent être décrits et contextualisés.

Dans ses finalités enfin. L'objectif de l'*open data* « c'est la recherche d'une deuxième vie, et d'une nouvelle utilité, pour tous les savoirs que crée l'État de par son activité quotidienne »[4]. Au-delà de leur fonction première, pour l'organisme qui les a produites, les archives, par l'entremise de celles et ceux qui

[1] *Vademecum sur l'ouverture et le partage des données publiques,* Secrétariat général pour la modernisation de l'action publique, Etalab, septembre 2013, p. 8
[2] *Ibid.*, p. 6
[3] *Ibid.*, p. 6
[4] http://www.henriverdier.com/2014/08/lopen-data-est-il-soluble-dans-la-big.html

les conservent connaissent de multiples vies. Les archives ont de nombreuses utilités. La loi sur les archives le rappelle, la conservation des archives est organisée tant « pour les besoins de la gestion et de la justification des droits des personnes physiques ou morales, publiques ou privées, que pour la documentation historique de la recherche »[1]. Aujourd'hui, on s'intéresse également à la valeur émotive des archives, au pouvoir créatif qu'elles suscitent et à leur exploitation par le milieu des arts.

Au-delà de cette convergence dans les principes, processus, moyens et finalités, quels liens entretiennent *open data* et archivage ? Comment mettre en évidence leurs complémentarités et envisager leur coexistence ?

La question est peut-être provocante mais l'*open data* est-il de nature à concurrencer les services d'archives, s'agissant de l'accès aux données publiques ? Les services d'archives ne sont plus en effet les seuls fournisseurs d'informations, au sens de sources primaires. Les plateformes de publication des données seront-elles nos futures salles de lecture virtuelles, pour ce qui est des données publiques ?

Une différence est de taille cependant. Elle tient au rapport qu'entretiennent l'*open data* et l'archivage avec le temps. L'*open data* s'inscrit dans le temps présent alors que l'archivage s'inscrit dans le temps long. Les données publiques publiées sont des données nécessairement actualisées, et donc mises à jour pour correspondre à la réalité du temps présent, c'est-à-dire du temps court de la consultation. Les données ou documents archivés ne sont pas modifiables et ne sont pas modifiés et c'est bien cette garantie qu'apporte le versement des archives dans un service d'archives. Dans ces conditions, la question de l'accès aux données que l'on pourrait qualifier de « périmées » du point de vue de l'*open data* se pose. De ce point de vue, on pourrait imaginer une continuité entre *open data* et archivage.

Pour autant, cette différence doit être mise au service de la complémentarité des deux processus car *open data* et archivage électronique poursuivent un objectif commun : obtenir des données de qualité. La qualité fait écho à la fiabilité : des données fiables, ce sont des données documentées, c'est-à-dire contextualisées. La réflexion autour d'un jeu de métadonnées commun pour la publication et l'archivage pourrait être un axe de coopération car on ne pourra pas demander aux producteurs des données de décrire deux fois leurs données et d'utiliser des formats de métadonnées différents !

Les données publiables ont la caractéristique d'être des données

[1] Article L211-2 du Code du patrimoine.

immédiatement accessibles parce que ne contenant pas d'informations à caractère personnel. En ce sens, elles correspondent, dans l'environnement papier, aux documents administratifs. L'*open data* s'inscrit en effet dans une filiation directe avec le principe de transparence administrative inscrit dans la loi sur l'accès aux documents administratifs du 17 juillet 1978. L'anonymisation est présentée comme le moyen de concilier protection de la vie privée et *open data*. La majorité des archives, quant à elles, comportent des données à caractère personnel. Le Code du patrimoine fixe le régime de communication des archives en définissant un seuil au-delà duquel les archives contenant des informations à caractère personnel deviennent accessibles, sans restriction. L'*open data* remet sur le devant de la scène les articulations nécessaires entre la règlementation sur l'accès aux documents administratifs, la protection de la vie privée et les archives. Et, au-delà de l'articulation, il s'agit de réaffirmer l'équilibre entre mémoire et oubli.

Les données publiées le sont par leurs producteurs qui identifient et sélectionnent les données publiables et les mettent directement à disposition des citoyens, sans médiation d'un tiers. *A contrario*, c'est l'archiviste qui analyse la valeur historique des archives et qui détermine lesquelles conserver. La mise en place, en France, de la fonction de *Chief data officier* (le titre officiel choisi par l'administration française est « administrateur général des données ») témoigne bien de cette nécessité du regard d'un tiers pour identifier les données publiables et organiser leur publication. De ce point de vue, les archivistes/ *records manager* auraient tout intérêt à mettre en avant leurs savoir-faire en termes de méthodes d'analyse de la valeur des données et de pratique des métadonnées normalisées.

On le voit, *open data* et archivage entretiennent un compagnonnage, dans les principes et les intentions. Réfléchir à leur articulation est indispensable sous peine de laisser naître des tensions. Cette articulation doit être pensée du point de vue théorique, législatif mais aussi pratique, au niveau des outils, interfaces et de l'urbanisation des systèmes d'information.

Open data **et archivage électronique : des convergences ?**

Comme nous avons pu le voir précédemment, la mise à disposition des données sur une plateforme *open data* et l'archivage peuvent répondre à des enjeux contraires, tout au moins difficiles à concilier. Le premier se situant dans l'immédiateté, le second dans le temps long, voire dans l'éternité.

Pour autant, les enjeux ne sont pas si éloignés l'un de l'autre, car ils nécessitent de pouvoir s'appuyer sur des données de qualité (pertinentes, authentiques, fiables, lisibles, intègres et documentées). Or, la qualité de la donnée dépend également de la capacité à la transmettre dans un espace-temps long. C'est ici que l'expertise de l'archiviste prend tout son sens : au cœur de la gestion de l'information, il occupe une des seules fonctions à avoir une vision transversale et exhaustive de la production documentaire, mais également, et peut-être surtout, du cycle de vie des documents, de la production à la réutilisation. Remarquez, et quel changement de paradigme, que l'on ne parle plus seulement de communication. Et si, comme ce n'est pas encore assez le cas, l'archiviste occupe également les fonctions de Personne responsable de l'accès aux documents administratifs et des questions relatives à la réutilisation des informations publiques (PRADA) et de Correspondant informatique et liberté (CIL), il peut devenir le pivot de l'*open data*. Ainsi, dans ce contexte de patrimonialisation des données – au sens de l'actif qu'il faut savoir conserver pour l'exploiter –, il doit prendre une part essentielle dans la gouvernance de l'information.

Le SAE ou l'aiguilleur des données vers l'*open data* ?

Le mouvement *open data* tend à changer le schéma de la production administrative : la donnée est produite afin d'assurer une mission mais doit également être envisagée comme pouvant servir à un autre usage (enjeu finalement assez proche de celui des archives). Dès lors, on peut penser que les systèmes d'information archivistiques, et notamment les SAE, ont un rôle fondamental à jouer dans la mise à disposition de données. En effet *:*

- le SAE collecte en direct de manière contrôlée et structurée des données (archives courantes et/ou définitives) ;

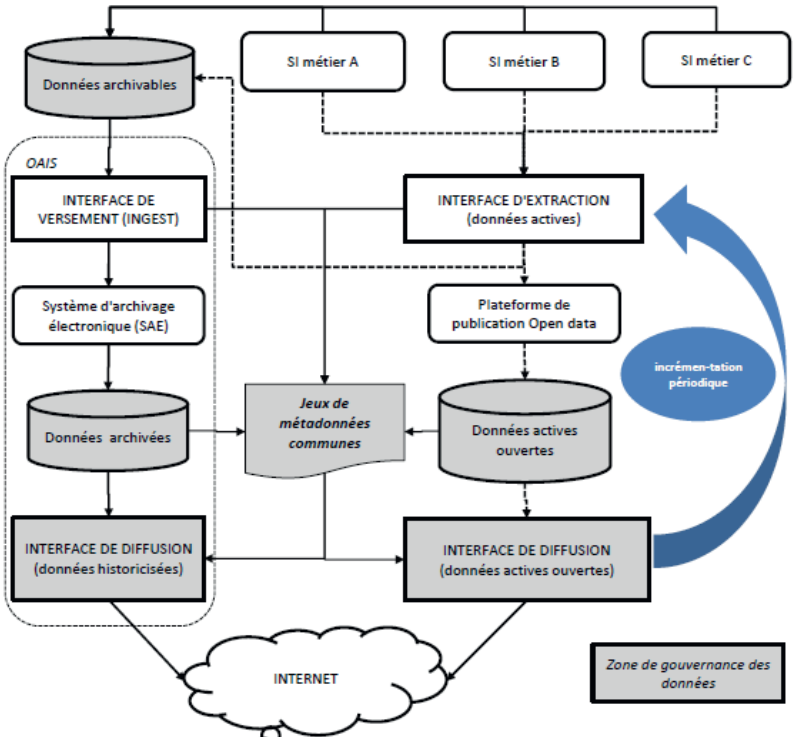
- le SAE fait converger différents paramètres : qualification, authenticité, intégrité et pérennité. Le SAE peut ainsi permettre de donner accès à des données courantes et intermédiaires mais également à des données historiques fiables. Si certaines données n'ont une valeur que dans l'instant présent (exemple : horaires de bus), d'autres ont une valeur à long terme car bénéficiant d'un historique (exemple : étude des flux de trafic routier).

- le SAE peut constituer un garant du contrôle de l'accès et la communication de données à caractère personnel.

Le SAE disposant alors de données structurées et qualifiées, donc de fait réutilisables, peut *via* un connecteur les verser sur une plateforme *open data*. Cela induit pour l'archiviste de déterminer en amont un nouveau critère : le caractère réutilisable ou non de la donnée.

Bien entendu, le SAE ne peut arbitrairement et automatiquement verser les données sur une plateforme *open data* sans la définition d'une politique éditoriale. C'est ici que la gouvernance des données prend tout son sens. De même, si le SAE peut être un aiguilleur de données, il ne peut en constituer le seul fournisseur. Par exemple, les données en temps réel ou celles provenant des systèmes d'information archivistiques (images, métadonnées de description des producteurs, etc.) peuvent alimenter directement une plateforme.

Le schéma ci-après illustre un processus mettant en parallèle le modèle OAIS de l'archivage électronique et la production de données en *open data* ainsi que leurs convergences possibles.



Systèmes d'information – Système d'archivage électronique - *Open data* © J.- D. Zeller, 2015

Une plateforme *open data* peut servir de fournisseur de données identifiées comme intéressantes à archiver et complétant le travail de collecte. La plateforme *open data* permet aussi, grâce à l'intervention des agents et des citoyens, une amélioration de la qualité des données produites. De ce fait, cela permet potentiellement au SAE d'être alimenté, de conserver et de mettre à disposition des données qualifiées et de qualité pour les chercheurs et réutilisateurs actuels et futurs. C'est ainsi que l'archivage électronique associé à l'*open data* peut être à l'origine de la création d'un nouvel écosystème de l'information vertueux.

L'archivage électronique prend alors un tout autre sens, son périmètre n'étant plus nécessairement restreint à la seule valeur probante de la donnée. Le bénéfice pour l'archiviste est ainsi de faire de la collecte d'archives électroniques un enjeu de diffusion et de réutilisation. Cela signifie qu'il doit s'affirmer comme partenaire auprès des promoteurs de l'*open data*.

Néanmoins, si l'on veut assurer cette convergence, il faudra s'attacher à construire des jeux de métadonnées communs, ce que les archivistes savent faire, et à inscrire ces processus dans le cadre plus vaste de la gouvernance de l'information.

Céline GUYON
Chargée de la politique de gestion électronique des documents et des archives
Conseil départemental de l'Aube
celine.guyon@aube.fr

Cyril LONGIN
Directeur
Archives municipales de Saint-Étienne
cyril.longin@saint-etienne.fr

Jean-Daniel ZELLER
Archiviste principal
Hôpitaux universitaires de Genève
Jean-Daniel.Zeller@hcuge.ch

ANNEXE

Biblio-webo-graphie

*Petite histoire de l'*open data, 2012 :
https://opendata.hauts-de-seine.net/comprendre/demarches/petite-histoire-de-lopen-data

CHIGNARD (Simon), *Open data*, FYP Edition, 2013:
http://www.fypeditions.com/open-data-comprendre-louverture-des-donnees-publiques/

CHIGNARD (Simon), *Une brève histoire de l'*open data, ParisTechReview, 23 mars 2013 :
http://www.paristechreview.com/2013/03/29/origines-open-data/

Secrétariat général pour la modernisation de l'action publique, *Vade-mecum sur l'ouverture et le partage des données publiques,* Etalab, septembre 2013 : http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/vademecum-ouverture.pdf

TROJETTE (Mohammed Adnène), *Ouverture des données publiques : les exceptions au principe de gratuité sont-elles toutes légitimes ?*, Rapport au Premier Ministre, La Documentation française, juillet 2013 :
http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/134000739.pdf