Une charte éthique pour le Big Data

Primavera de Filippi

Revue Documentaliste
Article de la rubrique Métier et compétence, 2 pages (p. 8-9)
10 Juin 2013

[outil] Des implications métiers, sans nul doute, pour ces obligations particulières qui doivent être respectées lors de l'utilisation de données provenant du Big Data, afin de garantir leur qualité, leur conservation et les conditions de leur réutilisation tant d'un point de vue technique, économique que légal. Une charte, outil pour nous y aider, vient d' être proposée.

Une charte éthique pour le Big Data

Après le Cloud computing, le Big Data est le buzz du moment. Il désigne des ensembles de données dont le volume, la diversité et la variabilité sont tels qu'ils s'avèrent difficiles à gérer avec des outils d'administration classiques.

Avec l'arrivée du Cloud computing et des réseaux sociaux, le développement de la téléphonie mobile, et le déploiement massif de capteurs ou de dispositifs intelligents (*smart devices*), on observe une croissance rapide et constante de la quantité de données disponibles sur le réseau. D'après l'International Data Corporation (IDC), on assistera, d'ici 2020, à une croissance moyenne de plus de 45 % par an en volume de données. Mais la plupart de ces données sont peu structurées et éparpillées dans le réseau en différents formats souvent non interopérables.

La création et la gestion de bases qui agrègent un grand nombre de données hétérogènes provenant de sources différentes est toutefois devenu une condition préalable à la fourniture de services innovants. L'interconnexion des données permet, en effet, d'obtenir de nouvelles informations et de réaliser des prédictions susceptibles de révolutionner les pratiques commerciales des grands opérateurs en ligne et de la société dans son ensemble. Ainsi, la production, l'utilisation, la gestion et la diffusion des données (personnelles, démographiques, scientifiques, ou autres) sont devenues des facteurs majeurs de compétitivité qui comportent des enjeux économiques toujours plus importants.

De la collecte à l'analyse sémantique des données, en passant par leur validation ou leur enrichissement, les données représentent aujourd'hui la matière première de l'économie numérique. Bien qu'elles demandent d'importants investissements, leur exploitation et leur réutilisation constituent une source considérable de revenus (actuels et futurs). De nombreux producteurs et fournisseurs de données ont réalisé que celles-ci avaient une valeur au-delà de l'objectif principal pour lequel elles avaient été originellement collectées. Une fois correctement structurées et organisées, elles peuvent être facilement réutilisées (par les producteurs initiaux ou par d'autres) pour des finalités de nature très différentes.

Or, bien qu'on ait développé de nouveaux outils pour faciliter le traitement de ces grandes masses de données, leur exploitation est limitée pour des raisons légales ou éthiques.

D'un point de vue juridique, elles ne peuvent être exploitées qu'après avoir identifié et négocié avec celui qui dispose des droits sur celles-ci. Lorsqu'il s'agit de données personnelles, la collecte, le traitement et la diffusion des données sur le réseau doivent se faire en conformité avec les dispositions de la loi Informatique et libertés de 1978, ce qui se traduit notamment par des obligations en matière de consentement (art. 7 de la loi) et de respect des finalités de traitement (art. 6).

D'un point de vue éthique, la difficulté majeure consiste à évaluer l'impact que l'exploitation et l'interconnexion d'une grande masse de données peut avoir sur le respect des droits et des libertés individuelles. Une autre difficulté consiste à assurer la conformité, la validité et l'intégrité des données fournies. Cela est dû notamment au manque de traçabilité des données, dont la provenance est parfois opaque et dont la qualité est souvent difficile à évaluer, comme dans le cas des données issues des communautés (*crowdsourcing*).

Charte Ethique Big & Data

En juin 2012, l'Association francophone de la communication parlée (AFCP), l'Association des professionnels pour l'économie numérique (Aproged), l'Association pour le traitement automatique des langues (ATALA) et Cap Digital ont constitué, en collaboration avec le CNRS, un groupe de travail "Ethique & Big Data" pour concevoir une charte visant à garantir la qualité, la traçabilité et la pérennité des données, tout en réduisant au maximum les risques juridiques liés à leur diffusion et à leur réutilisation. La Charte Ethique et Big Data est ainsi une construction conjointe réalisée dans le cadre d'un groupe de travail APROGED et Cap Digital. La diversité des participants académiques et industriels a favorisé l'expression de points de vues croisés qui se sont mutuellement enrichis pour aboutir à une Charte qui facilitera la création, la diffusion et la maintenance de données Big Data et participera ainsi à leur valorisation.

Ces aspects ont déjà été abordés à l'international[1]. En France, alors que l'on commence à peine à profiter des bénéfices et à se préoccuper des enjeux du Big Data, cette charte représente une initiative pionnière qui vise à examiner les questions éthiques soulevées par ce phénomène afin d'en explorer les répercussions sur les opérations et les décisions d'entreprises en ce qui concerne notamment la collection, le traitement et l'exploitation des grands jeux de données.

L'objectif est d'apporter des garanties en termes de transparence, usage, et rémunération à tous les acteurs impliqués dans la chaîne de valeur - producteurs, fournisseurs et utilisateurs des données - afin d'assurer le respect des lois et des éthiques professionnelles en créant un sentiment de confiance.

Cette charte fournit des indications aux fournisseurs de données, notamment pour les principaux éléments à prendre en compte ou à mentionner lors de la mise à disposition des données, quelles qu'en soient les modalités (payantes ou gratuites) et les finalités

[1] Kord Davis. Balancing Risk and Innovation. *Ethics of Big Data*. O'Reilly Media, 2012.

(commerciales ou non commerciales). Dans cette charte auto-administrée[2], ce sont les fournisseurs de données qui remplissent eux mêmes les informations concernant leurs données, et non un tiers auditeur ou certificateur (comme ce pourrait être le cas si une véritable certification était mise en place).

Quatre points clés

1. La description des données. C'est une étape fondamentale de l'analyse et du traitement des données. Il est nécessaire de préciser la nature des données, leur origine et la nature de leurs producteurs. Il est aussi important de donner des informations sur les supports et les modalités de fourniture et, si possible, d'apporter des références vers la source des données et tout document supplémentaire les décrivant plus en détail (par exemple, la description des modalités de collecte et de traitement, les liens vers des informations complémentaires, etc.).

2. La traçabilité. Elle permet d'identifier la source des données et de retracer leur évolution de leur création à leur diffusion. Il faut, pour ceci, donner des spécifications sur le processus de constitution des données ainsi que sur les différentes transformations qu'elles ont subies, notamment s'il s'agit de données primaires (produites directement par le fournisseur), consolidées (en agrégeant les données de différents fournisseurs) ou enrichies (à partir des données tierces). Pour chacune de ces catégories, il faut aussi donner des informations sur la typologie des contributeurs (entreprises, associations, individus, ou capteurs), la nature des relations contractuelles avec le fournisseur (salariés, contractants, stagiaires, volontaires) et, le cas échéant, sur les modalités et le montant des rémunérations.

Lorsque les données ont été modifiées, la nature et la typologie des données initiales, intermédiaires et finales doivent être clairement explicitées ainsi que le processus de transformation qui a été utilisé, en précisant notamment qui sont les acteurs de ce traitement et s'il s'agit d'un travail manuel, automatique ou semi-automatique.

Enfin, en ce qui concerne la qualité des données, il est important d'indiquer qu'un processus de validation a été appliqué (en interne ou par un organisme externe), les caractéristiques de ce processus (critères et fréquence de validation, métriques utilisées, etc.) et les résultats obtenus tant au niveau qualitatif que quantitatif.

3. La propriété intellectuelle. Les données - bien que non protégées par le régime du droit d'auteur - peuvent être protégées par le secret ou indirectement par le droit des producteurs de bases de données.

Le fournisseur doit donc garantir le respect des droits de chacun des producteurs de données en identifiant les droits portant sur ces données et les dispositions des licences sous lesquelles elles ont été mises à la disposition du public.

[2] Wiki de la charte http://wiki.ethique-big-data.org

Lorsque les données ont été produites par des tiers, le fournisseur doit en mentionner la source ainsi que toutes les restrictions (légales ou contractuelles) qui pourraient limiter leur réutilisation par des tiers. Une attention spécifique doit être accordée aux licences virales, de type *copyleft*[3], qui exigent que toute base de données dérivées soit mise à disposition du public sous les même conditions que la base de données originale.

Lorsque les données ont été enrichies par des tiers, le fournisseur doit indiquer les droits que chacun détient sur les données et les obligations qu'ils doivent respecter.

4. Réglementations spécifiques. Au delà des droits de propriété intellectuelle, certaines données (notamment les informations protégées par un secret, les données personnelles sensibles, les données financières, les données portant sur la santé, etc.) sont soumises à des réglementations spécifiques - sectorielle ou d'ordre public - telles que la loi Informatique et libertés.

Afin de collecter, utiliser ou diffuser ces données en toute légalité, les fournisseurs de données sont obligés de respecter ces réglementations. Ils doivent donc s'assurer que les données qu'ils fournissent ont été obtenues et mises à la disposition du public conformément au droit et aux réglementations applicables, tout en s'engageant à informer le public des contraintes ou obligations spécifiques pesant sur ces données.

La Charte Ethique & Big Data est déjà adoptée par l'Alliance Big Data (ADBS, Apeca, Aproged, Cap Digital, GFII, Institut Mines-Télécoms) et par des producteurs de données du secteur de la recherche.
Pré-annoncée lors du salon Documation de mars 2013, elle a été présentée lors des journées sur le corpus de référence du français du CNRS (http://www.ilf.cnrs.fr/spip.php?rubrique95), où elle a reçu un excellent accueil alors que plusieurs acteurs ont déjà déclaré leur intention de l'utiliser pour la diffusion de leurs ressources.

L'Alliance Big Data a profité du Forum du GFII pour réaliser une annonce officielle avec relais de communications médias. Une autre opération est prévue lors du colloque LILPA (http://corpus-lilpa.sciencesconf.org/) ainsi qu'en session plénière lors de TALN (http://www.taln2013.org/).

La charte est diffusée sous licence Creative Common et est disponible sur le wiki http://wiki.ethique-big-data.org/ et le portail de l'Alliance Big Data (HYPERLINK "http://www.alliancebigdata.com" www.alliancebigdata.com). Cette version est une première étape : les travaux continuent pour l'optimiser afin de répondre de façon plus complète aux différents besoins sectoriels et métiers.

Primavera De Filippi

CERSA / CNRS / Université Paris II

primavera.de-filippi@cersa.cnrs.fr

[3] http://fr.wikipedia.org/wiki/Copyleft