CHARTE ETHIQUE ET BIG DATA

FACILITER LA CREATION, L'ECHANGE ET LA DIFFUSION DES DONNEES

Version 14062013

ATALA
Association pour le Traitement Automatique des Langues

AFCP

ALLIANCE BIG DATA

proged

cap-digital Paris Region

gfii

Sommaire

PREFACE

La disponibilité des grandes masses de données (Big Data) permet d'en extraire des connaissances impossibles à appréhender autrement. Cela leur confère une importance stratégique et établit une barrière entre ceux qui peuvent y accéder et les autres. Dans l'objectif d'en garantir l'accès au plus grand nombre pour les besoins de la recherche, des initiatives ont été lancées au plan international pour partager ces données (Data Sharing). On peut mettre dans cette notion de partage une simple idée de distribution la plus ouverte possible, mais on peut aussi l'étendre à la production, la validation et l'enrichissement collaboratifs des données, et à leur utilisation pour développer et évaluer les technologies dans beaucoup de domaines. Cela passe par l'identification et la trace de l'utilisation de ces données, dans une approche qui doit être coordonnée et internationale pour pouvoir être effective. La myriadisation du travail parcellisé, ou crowdsourcing, peut être mise au service des activités de production et d'enrichissement des données. Elle apporte la possibilité d'établir un contact avec la « foule » internationale, et toute la force de travail qu'elle représente, mais soulève en même temps les problèmes éthiques d'une activité qui échappe aux règles habituelles du droit du travail. Consciente de ces enjeux et de l'urgence de les traiter, les rédacteurs se sont saisis ces questions et propose aux chercheurs et aux industriels cette charte afin d'encourager les aspects hautement positifs liés au Big Data, et décourager les effets potentiellement néfastes qui pourraient les limiter ou les inverser.

J. Mariani
Directeur de l'Institut des technologies Multilingues et Multimédias de l'Information (IMMI-CNRS)

La création, la maintenance, la diffusion et l'utilisation de données de toutes sortes est un enjeu économique majeur. Qu'il s'agisse de données démographiques, personnelles, de relevés de capteurs, de documents, thésaurus, ontologies. Ces bases de données sont essentielles à la création et la maintenance de nouveaux services. L'apparition du Cloud computing, de l'Open Data et du Big Data rendent ces questions particulièrement sensibles. Cependant, l'utilisation ou la réutilisation des données se heurtent trop souvent à des freins qui en empêchent l'exploitation optimale : provenance parfois opaque (en particulier dans les cas de " crowdsourcing "), traçabilité inexistante, protection intellectuelle incertaine, une qualité difficile à évaluer a priori. Dès lors, sécuriser la création de données est un facteur de compétitivité. Cette charte Ethique & Big Data se donne comme objectif de fournir des garanties concernant la maintenabilité des données, leur traçabilité, leur qualité, l'impact sur l'emploi, réduire le risque juridique. Cette charte vise à harmoniser les rapports entre producteurs, fournisseurs et utilisateurs de données sur le plan du respect des lois, de celui de l'éthique, et garantir la confiance dans les rapports entre l'ensemble des acteurs impliqués.

Alain Couillault,
APROGED,
Professeur associé Université de La Rochelle

Un travail à plusieurs voix

Cette charte a été conçue à l'initiative de l'APROGED, de l'ATALA, de l'AFCP et de CAP DIGITAL. Plusieurs associations et partenaires ont collaboré à sa rédaction et à sa diffusion.

Contenu de la charte

La Charte Ethique & Big data comprend quatre volets principaux qui concernent la description des données, la traçabilité, la propriété intellectuelle et les réglementations spécifiques. Pour chacun de ces volets, la charte considère la collecte et la fabrication de données, les processus d'enrichissement ou de transformation, et leur utilisation ou leur diffusion.

Comment utiliser cette charte ?

La Charte Ethique & Big data fournit une trame de description des données et sert de memorandum des points à décrire lorsque l'on met à disposition des données, que ce soit à usage commercial ou académique, payant ou gratuit, interne ou externe. Les éléments prévus dans la charte sont à remplir par le *fournisseur des données* qui les met à disposition et s'engage ainsi sur leur contenu.

Il arrive, fréquemment, qu'un jeu de données soit construit par rassemblement, enrichissement, altération d'un ou plusieurs jeux de données existants. Dans ces cas, il convient de remplir les éléments de la Charte pour le seul jeu de données auquel elle correspond, en y faisant référence, le cas échéant, aux chartes des jeux de données utilisés.

Licence

« Rédacteurs Gilles Adda, AFCP, CNRS-LIMSI, Christelle Ayache, Cap Digital, Alain Couillault, Apoliade, Aproged, Université de La Rochelle, Karën Fort, ATALA, Loria / LIPN, Pierre-Olivier Gibert, Digital Ethics, François Hanat, Cap Digital, Hugues de Mazancourt, Aproged, Eptica-Lingway.

Animateur du groupe de travail « Ethique et Big Data » organisé par l'Aproged : Alain Couillault,

Contributeurs : Daniel Bourcier, CNRS CERSA, Marie-Odile Charaudeau, Aproged, Primaveri de Filippi, CNRS CERSA, Olivier Itéanu, Aproged, Benoît Sagot, Aproged, INRIA/Paris VII, Joseph Mariani, CNRS Limsi/IMMI, Jamel Mostefa, ELRA/ELDA, Laurent PREVEL, Aproged. »

La charte est disponible en ligne à l'adresse http://wiki.ethique-big-data.org

Engagement

Par l'adhésion à la présente Charte, nous nous engageons, dans nos activités[1] relatives à l'accès, à l'extraction, à la réutilisation, à la fourniture de données dites Big Data, que ce soit dans le cadre d'un usage commercial ou académique, payant ou gratuit, à respecter les principes suivants :

Principe de traçabilité

Fournir les informations nécessaires et suffisantes pour que ces données soient utilisées en toute confiance, notamment concernant leur description, leurs processus de fabrication, de transformation, de vérification et de diffusion, les personnes intervenues dans ces processus ou la propriété intellectuelle.

Principe de respect des droits de propriété intellectuelle

Vérifier que les droits de propriété intellectuelle liés à la fourniture ou à la transformation des données sont respectés, et préciser la nature de la propriété intellectuelle sur les données que je fournis ou utilise.

Principe de respect des cadres légaux génériques et particuliers :

Identifier, notamment dans le cas de données personnelles, les règlements spécifiques à la nature des données traitées ou diffusées, et nous assurer que ces règlements sont respectés.

A cette fin, nous remplissons la présente Charte Ethique et Big Data et nous engageons sur les informations qu'elle contient.

A _____, le _____

[1] *La présente Charte concerne une activité générale non réglementée*

LES DONNEES

Les données

Nom du recueil de données :

Nom et coordonnées de l'institution ou de la personne responsable des données :

Personne(s) à contacter :

Responsable(s) de la charte :

Disponibilité des données (site Internet, CD-ROM...) :

Quelle est la nature des données fournies? Décrire les support, mode de fourniture (cf. support physique vs. flux d'informations) Si possible, indiquer précisément les références du document qui décrit les données fournies

TRAÇABILITE

Traçabilité

La notion de traçabilité couvre l'ensemble des aspects permettant de connaître le contenu d'une source de données, et de retracer le processus de fabrication,

Origine des données

S'agit-il :

- de données primaires (créées directement par le fournisseur),
- de données consolidées de différents fournisseurs
- de données construites à partir de données tierces (enrichissement) ?

Dans les deux derniers cas, fournir, pour chacune des sources la charte correspondante **ou** les coordonnées de l'organisation d'où viennent les données, ainsi que le contact permettant d'obtenir les informations afférentes, **ou** la mention explicite et argumentée que la charte ne s'applique pas.

Auteurs, processus de recrutement

Dans le cas de données primaires provenant de contributeurs humains, préciser

- la typologie des contributeurs
- la nature des relations contractuelles avec le fournisseur
- le mode de rémunération

Dans le cas d'utilisation de crowdsourcing, préciser :

- les critères de sélection des travailleurs,
- la ou les plateformes utilisées,
- le mode et le montant de la rémunération.

Si les données contiennent des données liées aux contributeurs humains, préciser :

- si un consentement a été demandé,
- si une trace matérielle existe de ce consentement.
- la nature de l'information fournie afin que le consentement soit éclairé,

Processus de fabrication ou de transformation des données :

A. Si les données dont l'origine a été spécifiée dans la section Origine des données ont subi une quelconque transformation:

- Décrire les processus de transformation.

B. pour les processus d'enrichissement de données,

- décrire la nature de l'enrichissement.
- Préciser pour chaque processus, s'il s'agit d'un travail manuel ou automatique

C. Dans le cas où un travail manuel est impliqué, indiquer :

- la typologie des intervenants,
- la nature des relations contractuelles,
- le mode de rémunération.

D. Dans le cas d'utilisation de crowdsourcing, préciser :

- les critères de sélection des travailleurs,
- la ou les plateformes utilisées,
- le mode et le montant de la rémunération.

E. Dans le cas où un outil informatique est impliqué, décrire :

- la nature et la fonction de l'outil,
- la nature de la propriété intellectuelle et la nature de la licence attachées à cet outil.

F. Dans le cas où les données contiennent des informations personnelles, préciser :

- les moyens permettant de s'assurer que la transformation est compatible avec le consentement décrit dans la section Auteurs, processus de recrutement,
- si une anonymisation a été effectuée, et la manière dont elle a été faite.

Processus de validation des données

G. Préciser si un processus de validation des données a été appliqué.

- Dans la négative, dire pourquoi un tel processus n'a pas été nécessaire

- Dans l'affirmative, décrire le processus de validation, et en particulier :
  - le pourcentage des données validées,
  - le mode de sélection des données validées,
  - si la validation a été faite en interne ou en externe,
    - si la validation a été externe, la nature de l'organisme de validation.
  - si la validation a été faite à l'aide d'outils automatiques, ou a nécessité une intervention humaine,
    - décrire la nature des outils,
    - préciser le profil des validateurs.
  - décrire la méthode de validation, et en particulier :
    - les critères de validation,
    - si ces critères impliquent l'utilisation de métriques, décrire ces métriques.
  - donner le résultat (qualitatif et quantitatif) de la validation,
  - s'il s'agit de données évolutives, indiquer :
    - si la validation est identique sur les données archivées, et les données nouvelles,
    - la fréquence de validation.

PROPRIETE INTELLECTUELLE

Propriété intellectuelle

Licence d'utilisation de(s) source(s)

H. En cas de réutilisation de données,

- décrire les restrictions légales ou contractuelles sur les données utilisées (par exemple, nature de la licence, la source doit-elle être citée? Etc.)

- La fourniture respecte-t-elle ces restrictions ? On veillera notamment à la viralité des licences affectées aux sources d'information. Par exemple, les sources sont-elles libres et ouvertes (OpenData...) ?

- Sont-elles soumises à une licence particulière ? à droit d'auteur ?

Droits du fournisseur sur les données

I. En cas d'utilisation de données tierces,

- le signataire de la charte a-t-il des obligations par rapport à ses fournisseurs? En particulier, l'origine des données (copyright) doit-elle être mentionnée ?

Altération de licence liée au traitement des données

J. En cas d'intervention d'un tiers sur les données (salarié, contractant, stagiaire...),

- préciser le cas échéant quels sont les droits de chacun sur les données (dans la mesure du possible, utiliser une licence pour préciser les droits et les obligations de chacun).

Licence d'utilisation

- Préciser la (ou les) licence(s) attachée(s) aux données fournies. (on veillera à ce que la licence précise s'il existe des restrictions quant à la rediffusion de ces résultats).

REGLEMENTATIONS SPECIFIQUES

Réglementations spécifiques

Certaines données peuvent être soumises à des réglementations d'ordre public qui s'imposent pour des raisons impératives de protection, de sécurité ou de moralité. Les fournisseurs ne peuvent y déroger. Le non-respect de ces réglementations peut donner lieu à des sanctions pénales ou prononcées par des autorités administratives indépendantes (CNIL, AMF, Autorité de la Concurrence).

Le respect de ces réglementations est donc une des conditions de la légalité de l'utilisation ou la réutilisation des données.

- Préciser si la nature des données fournies ressort d'une ou plusieurs réglementations spécifiques. Si oui, préciser la ou lesquelles.

- le fournisseur respecte-t-il ces réglementations ?

Il est de la responsabilité du fournisseur de rechercher les réglementations applicables.

Pour information, il existe des réglementations d'ordre public qui visent explicitement les données :

- Loi informatique et libertés relatives aux données personnelles
- Sur le site de la CNIL
- Droits des producteurs de bases de données (LIVRE III - Titre IV du Code de la Propriété intellectuelle)

Par ailleurs, suivant les secteurs d'activité, des réglementations spécifiques peuvent nécessiter de modifier les conditions de collecte d'utilisation et de réutilisation des données. Préalablement, à la mise en œuvre du traitement, une recherche sur les réglementations applicables s'impose. En particulier, les activités traitant des données personnelles, financières, de santé ou couvertes par un secret doivent faire l'objet d'une vigilance particulière.

Le correspondant informatique et libertés (CIL), présent dans de nombreuses entreprises, administrations ou collectivités locales est par exemple à même de renseigner ou d'instruire ce type de demandes.