# 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 proposent 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,*

# 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



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

# 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 :**

**Provenance de la source des données :**

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

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

**Pour quelle utilisation ces données ont-elles été initialement recueillies?**

# 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.

## Date ou période de collecte ou de constitution

## Nature des données

S'agit-il, au moins pour partie :

- de données personnelles (au sens de la CNIL)
- de données personnelles sensibles (au sens de la CNIL)
- de données collectées dans un pays tiers à l'UE

  Dans le derniers cas, préciser ce ou ces pays

## Auteurs, processus de recrutement

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

- la typologie des contributeurs (étudiants, salariés, patients, etc)
- 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 les contributeurs étaient à même de donner leur consentement
- la nature et la forme (texte ou orale) de l'information fournie pour obtenir ce consentement (par exemple consentement libre et éclairé au sens du Code de la Santé Publique)
- si une trace matérielle existe de ce consentement.

## Processus de fabrication ou de transformation des données :

- Date ou période de transformation

### 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
- Une politique de protection et de sécurité des données a-t-elle mise en œuvre ? La décrire le cas échéant

### B. Pour les processus d'enrichissement de données,

- décrire la nature et le type d'enrichissement.
- Préciser pour chaque processus, s'il s'agit d'un travail manuel ou automatique
- Indiquer pour chaque dispositif d'enrichissement, la durée de réalisation

### 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 de l'outil et son rôle dans le processus de transformation,
- 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,
    - Préciser 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 réalisée en interne ou en externe,
    - si la validation a été interne, le statut et la fonction de la personne.
    - 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, citation de la source, nature du consentement 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 ?

  Quelles sont les finalités du traitement de ces données tierces ?

## 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 l'utilisation ou 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, ANSM).

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 ?

Y a-t-il d'autres points liés à l'éthique qui, selon vous, doivent être mentionnés :

Si oui, lesquelles :

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.