1 Introduction

Data banks – and therefore their creation, distribution, maintenance, and use – are key contributors to economic competitiveness. They provide the raw material for the creation of new innovative services, be it demographic or personal data, documents, semantic web ontologies, thesauri, sensors inputs... Moreover, new trends such as Cloud Computing, Open and Big Data have put forward data banks of all kinds. Nevertheless, the optimal use – and reuse – of data is too often refrained due to external constraints: where does the data come from (especially when the data is created from crowdsourcing)? How can it be traced back? How is its quality ascertained? Which IP rights are attached to the data? Clearing these issues for each data set turns out to be a lever for competitiveness.

The goal of the *Ethics & Big Data Charter* is to bring enhanced security regarding the maintenance, traceability, quality, impact on employment or legal risks. It aims at smoothing the relation between data banks creators, distributors and users, both on the legal and ethical standpoints.

2 How to use this Charter

The *Ethics & Big Data Charter* provides a checklist for describing a data set. It should accompany each data set provided, be it in a commercial or academic context, free or for charge. The check list is filled in and signed by the provider, who thereby commits to its content.

3 Licence

- writers: 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.

Task force leader: Alain Couillault,

- Contributors: 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.
  English translation reviewed by Kevin Bretonnel Cohen, University of Colorado at Boulder.

4 The *Ethics and Big Data Charter*

4.1 Describing the data set

4.1.1 Name of the dataset

4.1.2 Name and coordoniates of the institution or person in charge of the dataset

4.1.3 Contact name

4.1.4 Availability

What type(s) of data does the data set contain? Describe the type of media (e.g. physical media, data flow, ...). If possible, provide the reference to a document which precisely describes the content of the data set.

4.2 Traceability

Traceability refers to all information pertaining to the history of the data which makes it possible to trace back its processing from creation to diffusion.

4.2.1 Origin of the data

Is the data set composed of:

- ☐ primary data (created first-hand by the provider),
- ☐ consolidated data (from one or various providers),
- ☐ enriched data (built from/over third-party data)?

In the case of consolidated or enriched data, provide either:

- the relevant *Charter*,
- or the references and contact name of the organization which provided the data,
- or explain the reason that the *charter* does not apply to the data.

4.3 Authors, recruitment

In the case of primary data created by/with human contributors:

- What are their profiles (skills, background, ...)?
- What is the nature of the contractual relationship with the provider?
- How are they remunerated (by the hour, by the task, ...)?

In the case of crowdsourcing:

- By which criteria were the contributors selected?

- Which platforms were used?
- What was the remuneration model and amount?

If the data set contains personal or human-related data:

- Were the individuals asked for their consent?
- What information was provided to them to ensure their informed consent?
- How was the consent obtained?

4.4 Creation and transformation of the data

If the data – as mentioned in section 4.2.1 (*Origin*) – has been altered:

- Describe the process put in place for altering the data
  - If relevant, describe what information has been added to the original data
- Was the alteration performed manually or by means of an automatic process?
  - If the alteration was (entirely or partly) done manually by contributors:
    * What were/are their profiles (skills, background, ...)?
    * What was/is the nature of the contractual relationship with the provider?
    * How were/are they remunerated (by the hour, by the task, ...)?

    In the case of crowdsourcing:
    * By which criteria were the contributors selected?
    * Which platforms were used?
    * What was the remuneration model and amount?
  - If a computer program was used:
    * Which task did the program perform (describe its input and output)?

    - Describe the intellectual property rights and license attached to the program:

- if the data set contains personal data:
    - How did the provider ensure that the alteration complies with the consent mentioned in section 4.3 (*Author and recruiting*)?
    - Has the data been anonymized? If yes, how was the anonymization performed?

4.5 Data validation

Was the data validated through a formal procedure?

- If not, explain why
- If yes, describe the validation procedure:
    - What portion of the data was validated?
    - Which data was validated?
    - Was the validation performed in-house or was a third party called for?
        - if relevant, what type of third party was involved in the validation process?
    - Validation means:
        - Which type of computer program was used to perform or assist the validation?
        - What are the profiles (skills, background, ...) of the validators?
    - Describe the process used for the validation of data:
        - Which validation criteria were applied?
        - Describe the types of metrics (i.e. score-based measure) used
    - Provide the qualitative and quantitative results of the validation
    - In case of data flow or data updates:
        - Is the validation process applied similarly to the original and the new data?
        - How often is the validation performed?

5 Intellectual property

The issues mentioned in this section are usually addressed by the license agreement. If so, it is sufficient to refer to the license agreement attached to the data set.

5.1 License agreement for the original data

If third party data is used, describe the legal or contractual restrictions which apply (for example, what license applies? Must the source be cited?...)

Does the provider comply with these restrictions? (attention should be paid to viral effects of certain license agreements)?

5.2 License alteration

If an individual (employee, contractor, trainee...) or third party is involved in the collection, creation or modification of the data, does this imply changes in the licensing scheme? (For example, does this individual or third party gain copyright or citation rights to the resulting data set?) When possible, refer to a license to describe the rights and obligations of each party involved.

5.3 License

Under which license is the data set provided?

6 Specific laws

Data sets may fall under specific laws depending on the nature of their content on the basis of protection, security or moral policies which providers must comply with. This is particularly true for financial, health or personal data.

- Does the information provided fall under specific laws or limitations?
  - If yes, which laws or limitations apply?
  - Does the provider comply with these laws or limitations?