Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter

Alain Couillault[*], **Karën Fort**[†], **Gilles Adda**[◇,⋆], **Hugues de Mazancourt**[‡]
[*] Université de La Rochelle/L3I, Av. Michel Crépeau, 17042 La Rochelle Cedex 01, France, alain.couillault@univ-lr.fr
[†]Université de Lorraine/LORIA, 54500 Vandœuvre-lès-Nancy, France, karen.fort@loria.fr
[◇] LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France, gilles.adda@limsi.fr
[⋆] IMMI-CNRS, rue John von Neumann, Orsay , France, adda@immi-labs.org
[‡] Eptica-Lingway, hugues.de-mazancourt@eptica.com

Abstract

The authors have written the *Ethics and Big Data Charter* in collaboration with various agencies, private bodies and associations. This Charter aims at describing any large or complex resources, and in particular language resources, from a legal and ethical viewpoint and ensuring the transparency of the process of creating and distributing such resources. We propose in this article an analysis of the documentation coverage of the most frequently mentioned language resources with regards to the Charter, in order to show the benefit it offers.

Keywords: language resources, ethics, documentation, big data.

1. Introduction

To adopt an ethical behavior in developing, funding, using or promoting language resources is first and above all a matter of choice: for the provider, deciding which approach to adopt – crowdsourcing or not –, or which platform to request on, or the level of remuneration of the workers; for the funding agency, choosing which projects to fund; for the users, choosing which resources to use or acquire. These choices have to be learned ones.

We designed the *Ethics and Big Data Charter* (Couillault and Fort, 2013) in collaboration with representatives from interest groups, private companies and academic organizations, including ELDA[1], the French CNRS[2], ATALA[3], AFCP[4] and APROGED[5]. The purpose of this Charter is to provide resources developers with a precise framework to document their resources and ensure their traceability and transparency.

This paper introduces the *Ethics and Big Data Charter* and evaluates the benefits it brings by comparing the information the Charter requires with that available for the major existing language resources.

We first present the *Ethics and Big Data Charter*, then we detail the methodology we used to compare existing language resources documentation with the content of the Charter and we present and discuss the results we obtained.

[1]Evaluations and Language resources Distribution Agency, `http://www.elda.org/`

[2]Centre National de la Recherche Scientifique/National agency for scientific research `http://www.cnrs.fr/`

[3]Association pour le Traitement Automatique des Langues/Natural Language Processing Association `http://www.atala.org`

[4]Association Française de Communication Parlée/French spoken communication association, `http://www.afcp-parole.org`

[5]Association de la Maîtrise et de la Valorisation des contenus/Association for mastering and empowering content, `http://www.aproged.org`

2. The *Ethics and Big Data Charter*

2.1. A Guide for Good Practice

The *Ethics and Big Data Charter* stands as a good practice guide for documenting language resources in terms of traceability, copyrights and ethics. It is provided as a self administered questionnaire to be completed by the language resources owners or distributors. When used beforehand to building the language resources, the *Ethics and Big Data Charter* can serve as a check list for the project leader. Examples of the questions are provided in section 3.3.

2.2. From Language Resources to Big Data

In the process of designing the Charter, it soon appeared that the issues raised for language resources apply to a larger range of data sets, which can be described as Big Data. Indeed, Big Data are characterized not only by their volume, variety and speed of creation, but also by their complexity, which characterizes even small sets of language resources. Reversely, the work conducted for language resources can be generalized to and benefit to Big Data sets.

2.3. Contents of the Charter

The *Ethics and Big Data Charter* is split into three major sections:

1. Traceability
2. Intellectual property
3. Specific legislation

They are preceded by a short identification section containing the name of the resource, the contact and responsible persons and a short description of the language resources.

2.3.1. Traceability

Traceability is key to our purpose of putting forward ethical issues. The traceability part of the Charter allows to give specific details about the relationship between the resource provider and the workers involved in developing the

resource, including legal bounding, workers skills and selection criteria.
Specific focus is put on personal data, i.e. data, like voice or video recordings, which can provide means to identify a person directly or indirectly. The Charter requires to describe if and how the data is de-identified, and if and how the contributors were informed of the purpose of the data collection. For example, the Charter was used in a research project for which the author collected and annotated recorded data from patients who underwent thyroidectomy (Fauth et al., 2013).
Quality assurance is another major aspect of traceability addressed by the Charter, as it requires to document the quality assurance strategy, so that the user of the data set is fully informed on the level of quality s/he can expect: what QA procedure the data were passed through? what portion of the data has been evaluated? What are the actual metrics that were used and the obtained results?

2.3.2. License and Copyright

Thanks to a great deal of effort accomplished in the definition of – mainly open source – licensing schemes, it has become common practice to attach a license to a data set.
The License and Copyright section of the Charter goes beyond this and puts the focus on questions which may be disregarded, like ensuring that the legal or moral copyrights of the persons who worked on compiling, enriching or transforming the data are respected.
As an example, we saw to it that all the writers of the *Ethics and Big Data Charter* are mentioned in the license citation. Also, the Charter reminds data collectors and distributors that they should check whether they comply with any third party data license they may use. As an example, the *Ethics and Big Data Charter* related to the TCOF-POS annotated corpus (Benzitoun et al., 2012) refers to the TCOF corpus on which it is built.

2.3.3. Specific Legal Requirements

A third section of the *Ethics and Big Data Charter* deals with legal requirements that may arise from certain properties of the data set.
For example, a country may have issued specific laws regarding the storage, use and/or dissemination of personal data. The Charter serves as a reminder for checking if such requirements exist.

2.4. Availability

The *Ethics and Big Data Charter* is available on-line.[6] The Web site is currently in French, and an English translation of a charter is also available[7]. Cap Digital, the French projects screening agency which participated in the creation of the *Ethics and Big Data Charter*, has created an on-line form to help companies proposing projects to fill in the Charter corresponding to their data sets.[8]

[6] `http://wiki.ethique-big-data.org`

[7] `http://wiki.ethique-big-data.org/chartes/charteethiqueenV2.pdf`

[8] The form is available at: `http://form.jotformeu.com/form/32473349455359?`

Examples of Charters are also provided, including one for TCOF-POS (Benzitoun et al., 2012), one for a corpus of E-mail messages created during the EU/Feder-funded GramLab project[9] (Couillault et al., 2013) and one for a medical dataset (Fauth et al., 2013). Some of these corpora raise non-trivial privacy issues that the *Ethics and Big Data Charter* allows to deal with.

3. Evaluating Language Resources Documentation

The purpose of this evaluation is to check whether the major language resources used today in Natural Language Processing (NLP) provide easy access to all the information covered by the *Ethics and Big Data Charter*. To achieve this, we first selected a set of resources, then applied the *Ethics and Big Data Charter* to these resources and computed the proportion of answers to each question.

3.1. Selecting Resources

We chose to rely on the LRE map (Calzolari et al., 2012) and focused on the resources which have the highest *Impact Factors*, i.e. which have been the most frequently cited among LREC 2010, LREC 2012 and COLING papers.
The resources we considered are the ones which are cited at least five times for the *data* and *tools* categories and more than once for the *evaluation* and *Metaresources* categories. They range from the Princeton WordNet (Fellbaum, 1998) to Europarl (Koehn, 2005), for the most cited resources and include resources in French (L*efff* (Sagot, 2010)), Japanese (Doshisha eye-gaze dialogue data (Jokinen et al., 2010)), as well as both oral and written resources.

3.2. Collecting and Evaluating the Documentation

For each of these resources, we considered only the information provided on the Web site where the resource is made available (the one mentioned on the LRE map Web site), along with one or two major related articles, when available, so as to ensure that we only took into account information that is easy to find by someone willing to use the language resources.
We created a grid with YES/NO questions and filled it by hand[10]. Our goal was to check whether the questions raised by the *Ethics and Big Data Charter* are documented for the considered language resource. For example, for the section dedicated to crowdsourcing platforms, the grid contains the following questions:

- Was a crowdsourcing platform used in the process of building or transforming the data? If yes:
  - Are the criteria used to screen the contributors provided?
  - Is the name of the platform(s) provided?
  - Are the wages provided?

[9] See: `http://www.gramlab.org`.

[10] The grid is available at `http://wiki.ethique-big-data.org/papers/formulaire.ods`

For each information item in the grid, we assigned 1 when we could find the information, and 0 otherwise. We shared among ourselves the work of filling this grid for each resource and organized intermediate meetings to ensure consistency.

3.3. Results and Discussion

In the tables below, the first column contains the question as it appears in the *Ethics and Big Data Charter*, the figures indicate the percentage of resources for which the information is available whether from the Web site where they are distributed or the major article which describes them.

3.3.1. Whom to Contact?

All the resources we evaluated provide basic information such as whom to contact, what institute originated the data, how to get the data, and what type of data is made available.

Question	% replies
Name and contact details of the person responsible for the data set	92
Person to contact	100
Data availability (Web site, CD-ROM...)	88

3.3.2. What has been done?

Similarly, the work accomplished to create or transform the language resource is most often very well documented, and so is the origin of the data, including when third party data were used (i.e. what we called *secondary* or *consolidated* data).

Question	% replies
What is the nature of the provided data (primary, secondary or consolidated)?	92
Describe how the data was transformed	81
(if the data was enriched) Describe what information was added	90
(if a computer program was used to transform the data) Describe the purpose of the computer program	84

3.3.3. Under which License?

The license under which the data is made available is documented for half of the resources, this despite the availability of a wide choice of well advertised licensing schemes such as the Creative Commons[11] licensing scheme or the GNU licenses[12].

In addition, the language resources which are built, at least partially, from other resources, rarely mention the potential related legal limitations, despite the fact that some of the external resources are provided by private bodies which may have put restrictive licenses on the resources. Hence, the risk is either that users of the resources may unknowingly infringe copyrights or simply turn away from the resource.

Question	% replies
Under which licence are the data provided?	50
Do the data fall under specific licensing constraints	15
(if the data set includes third party data:) Describe the implied legal restrictions	35
Are there requirements with regards to third party data?	47

3.3.4. Who Worked?

While most of the language resources involve manual work, the profiles of the contributors, their legal relationships or remuneration levels are rarely described. This is a concern for two reasons. First, it prevents resource users to have access to the full traceability on the data and may raise issues regarding the contributors potential copyrights on the data or even regarding their skills: have they been trained for the specific tasks? Are they experts on the specific domain of the data or at the required task to annotate them? This information would certainly be valuable to track potential bias or understand where the consensus on quality assurance comes from.

In addition, it would be useful for the recruitment of workers for further language resource projects.

The figures provided below refer to the 71% of the resources which involve human resources.

Question	% replies
Skills of the workers?	39
Type of contract they work under	0
Type of remuneration (salary, subcontracting...)?	0

3.3.5. Crowdsourcing

The figures below are relative to the 11% of the resources which have been described as using crowdsourcing.

Question	% replies
Upon which criteria were the wrokers selected?	33
Which crowdsourcing platform was used?	67
What was the workers remuneration	33

3.3.6. Quality Assurance?

We found a surprising lack of information regarding the quality assurance process of the provided data. We consider that information on quality is a major aspect of traceability and description of the data. This information is often crucial when the data is used for the training or evaluation of

[11] http://creativecommons.fr/

[12] https://www.gnu.org/licenses/licenses.html

a specific tool or, more generally, for any type of research which would rely on the data.

Question	% replies
Was a quality assurance procedure applied to the language resource?	38
(if not) why was no quality assurance procedure applied?	0
(if yes) describe the quality assurance procedure applied	47[13]
(if yes) provide the qualitive and quantitative results	53[14]

4. Conclusion

The present study shows that even the most widely referred to language resources listed in the LRE map present documentation lacks, in particular regarding the persons who produced the work and which external resources have been used.

The reality we are facing now in the call for projects of all the national and international funding research agencies is a growing interest for Big Data, and a foreseeable burst in the number of Big Data related projects, with the use, for instance, of personal data from Twitter or Facebook, the development of crowdsourcing and so on. But we also know that for these data, uncertainties about privacy, the way they have been acquired, who has really worked, is likely to be of an order of magnitude more important than for the present resources listed in the LRE Map.

Language resources are viral, in that most of the language resources (75%) we examined are built from other resources, which may in turn be built from previous resources: for example, the *Prague Czech English Dependency Treebank* (Hajič et al., 2012) is partially built from the *Penn Treebank*(Marcus et al., 1993) , which is an annotated version of a first hand corpus. This implies that lack of information, especially regarding every aspects of quality assurance and of copyright issues, may reduce the usability of the resources for further work. It also means that the investment granted for a well designed and documented language resources can leverage the creation of further language resources.

For these reasons, we think it is crucial to gather all the initiatives such as the *Ethics and Big Data Charter* which aims at promoting ethics and traceability in resources, in order to propose, at the international level, a way to limit the risks for all the actors (funding agencies, research laboratories, private companies) of the data added value chain, regarding the use of Big Data and bring the benefits of the *Ethics and Big Data Charter* to current projects such as the language resources unique reference number (Choukri et al., 2012) or the METASHARE schema (Gavrilidou et al., 2012).

5. Acknowledgements

The *Ethics and Big Data Charter* is provided under the Creative Common licence BY 3.0 FR. We wish to thank the other authors of the Charter: Christelle Ayache and François Hanat, from Cap Digital and Pierre-Olivier Gibert, from Digital Ethics. We also thank the contributors: Danièle Bourcier and Primavera de Filippi (CNRS CERSA), Joseph Mariani (CNRS Limsi/IMMI), Marie-Odile Charaudeau, Olivier Itéanu and Laurent Prevel (Aproged), Benoît Sagot (ATALA, INRIA/Paris VII) and Jamel Mostefa (ELRA/ELDA).

This work has been partially funded by the Tourinflux Project [15], within the framework of the French government funded *Investissement d'avenir* programme.

Finally, we thank Joseph Mariani (LIMSI-CNRS, IMMI) and Gil Francopoulo (Tagmatica) for their help with the LRE map.

6. References

Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Proc. of Traitement Automatique des Langues Naturelles (TALN)*, pages 99–112, Grenoble, France, June.

Calzolari, N., Gratta, R. D., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., and Soria, C. (2012). The LRE Map. harmonising community descriptions of resources. In *Proc. of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May.

Choukri, K., Arranz, V., Hamon, O., and Park, J. (2012). Using the international standard language resource number: Practical and technical aspects. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Couillault, A. and Fort, K. (2013). Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Proc. of the international colloquium Corpus et Outils en Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasburg, France, July. 4 pages.

Couillault, A., Vinckx, A., de Mazancourt, H., Grandry, F., and Recourcé, G. (2013). Use case eptica/lingway : identification d'amorces de reprise. Technical report, GramLab Project ; EU/Feder funded project.

Fauth, C., Vaxelaire, B., Rodier, J.-F., and Sock, R. (2013). Corpus en parole pathologique, intérêts et enjeux : l'exemple d'un corpus enregistré à partir de patients thyroïdectomisés. In *Proc. of the international colloquium Corpus et Outils en Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, Strasburg, France, July.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., and Mapelli, V. (2012). The meta-share metadata schema for the description of

[15] http://www.tourinflux.com/

language resources. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpcnek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing prague czech-english dependency treebank 2.0. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Jokinen, K., Nishida, M., and Yamamoto, S. (2010). On eye-gaze and turn-taking. In *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction*, EGIHMI '10, pages 118–123, New York, NY, USA. ACM.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proc. of the international conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.