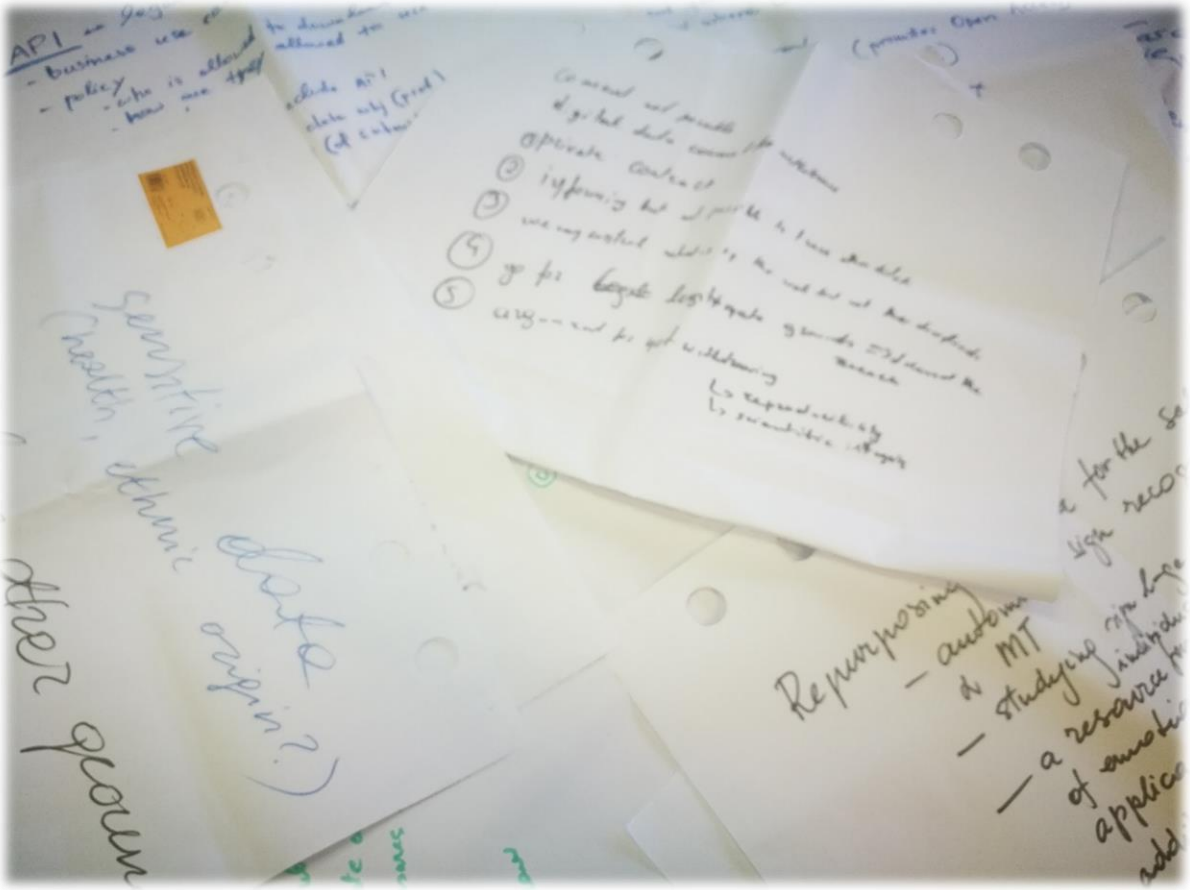
Hacking the GDPR
to Conduct Research with Language Resources
in Digital Humanities and Social Sciences

Final workshop report



Contents

Workshop program
(with links to presentation slides)

Date: 7 December 2018
Place: Vilnius, Lithuania
Address: Crowne Plaza hotel (http://www.cpvilnius.com/), Mk Ciurlionio Street 84, Vilnius

Overview

The workshop, organised by the CLARIN Legal and Ethical Issues Committee (CLIC) and hosted by CLARIN-LT, aims to bring together legal experts and researchers from the Digital Humanities and Social Sciences disciplines working with Language Resources (LRs) in order to exchange views and explore ways of creating and using LRs under the GDPR regime. Furthermore, the proposal for establishing a Code of Conduct will be laid out.

The workshop is organised in two sessions:

- an introductory session with short presentations related to the topic aiming to establish a common ground of knowledge among the participants
- a "hackathon" session where the participants will collaborate on resolving three Use Cases that present problems vis-a-vis the GDPR; participants will have the opportunity to discuss the legal and technological measures that can be used to ensure legal access to and processing of personal data, comparing and contrasting the solutions in order to select among them the ones that best suit each case, thus familiarising themselves with GDPR concepts.

IMPORTANT NOTICE: The workshop is meant for academic discussion and nothing said during the workshop can be construed as legal advice.

Detailed program

Room: Coral B, 2nd floor

Session 1: Introduction to GDPR and related issues

09:00 - 09:05	Opening and welcome *Jurgita Vaičenonienė, CLARIN-LT*
09:05 - 09:20	Hacking the use of personal data on consent (slides) *Alexandros Noussias, CLARIN:EL*
09:20 - 09:35	Use of personal data without consent (slides) *Aleksei Kelli, CLARIN-ET*
09:35 - 09:50	General framework for research use and proposal for a Code of Conduct (slides) *Pawel Kamocki & Erik Ketzan, Germany*
9:50 - 10:05	Overview of protection measures (slides) *Krister Lindén, FIN-CLARIN*
10:05 - 10:15	Questions/Discussion
10:15 - 10:45	*Coffee break*

Session 2: Hackathon & Short discussion

10:45 - 11:00	Introduction to hackathon (slides) *Krister Lindén, FIN-CLARIN*
11:00 - 12:30	Use Case 1: Sign Language Dictionary* (slides) Presented by *Krister Lindén* and discussed by *ALL*
12:30 - 13:30	*Lunch*
13:30 - 15.10	Use Case 2: DH Course registry * (slides) Presented by *Vanessa Hannesschläger* and discussed by *ALL*
14:20 - 15:10	Use Case 3: Web corpus * (slides) Presented by *Serge Sharoff* and discussed by *ALL*
15.10 - 15.30	Small hackathon Use cases ** • Anna Maria Bruzzone's speech archive *(Silvia Calamai)* (slides) • Slovenian use case (*Mateja Jemec Tomazin*) (slides) • PFC corpus *(George Christodoulides)* (slides) • Textual input from vulnerable people (*Ineke Schuurman)* (slides) • ReadLet project (*Riccardo Del Gratta*) (slides)
15:30 - 15:45	*Coffee break*
15:45 - 16:00	Wrap-up conclusions & way forward

* For each use case, the following outline will be followed:

1. Presentation of the use case (5-10 min)
2. Split into three groups: (10-15 min)
    a. legal experts – initial analysis of how to formulate the legal basis for collecting and using the research data (with possible national twists)
    b. researchers (humanists/social scientists) – what research questions would linguists/humanities scholars wish/need to use the data for
    c. researchers (technologists) – what technical protection measures could be used without compromising the quality of the data or unnecessarily complicating access to the data
3. Each group present their initial solution (3 x 5 min)
4. Split into three groups with at least one representative from each of the previous groups to discuss and modify the solutions to take into consideration the additional information that was presented (15 min)
5. The groups present their modifications and additional solutions and the other groups comment (3 x 5-10 min)

** For each use case in the small hackathon, a short presentation (2-3 minutes) will be followed by a plenary discussion.

IMPORTANT NOTICE: The workshop is meant for academic discussion and nothing said during the workshop can be construed as legal advice.

Workshop notes & conclusions

IMPORTANT NOTICE: All following notes are meant as part of an academic discussion and nothing noted here or said during the workshop can be construed as legal advice.

Introductory session (cf. slides in the program)

- [by Krister] There is a need for a new license, that would cater for resources that contain sensitive data that are not (and should not be) anonymised in order not to hinder research, but which should on the other hand be protected in order not to put the data subjects at risk. This license should be easy to read and understand (also for non lawyers, i.e. researchers and data subjects).
- The Code of Conduct should be implemented in the license.

Use case 1 (Sign Language Dictionary)

- Infrastructure's point of view:
  - need to consider who is the *data owner*, the *data controller* and the *data user*
  - legal responsibilities at all levels of *data depositing*, *using*, *re-purposing*, *disseminating*
  - Consent forms:
    - they need to be documented and stored; permanently? should they also be part of the data package when downloaded? or part of the metadata?
    - important to add in the new consent forms a list of potential risks
- A number of things to consider for public licences and web distribution
  - digital data cannot easily be withdrawn; even if withdrawn, impossible to trace all distributions that have already been downloaded
  - public licences are incompatible with purpose limitation => Code of Conduct can be helpful for this
  - Data distributed with public licences cannot be withdrawn => (a) inform participants of the intention to distribute data under open licence and (b) mark this clearly at the distribution point
  - Public licenses such as **Creative Commons** focus on copyright and are limited to the use of data. They do not specify anything about personal data; thus, they are **incompatible with restrictions on the use of datasets due to personal data contained**.
- Points to consider with respect to the resource:
  - does it include special categories of Personal Data (PD)? seems so (ethnic origin, health)
- Better treat the datasets as different and handle them differently
  - As a general comment, for the "sharable" part, using CC-BY(-SA) should be encouraged to promote Open Access (i.e. avoid NC if possible)
  - For the material without consent, it seems that CC-BY is not possible ⇒ better to share "only for research"; with restricted access to the material, we can control who downloads what; yet, again we cannot trace if they share with other users

- Preferable to go for "research on legitimate grounds"; but in order to support this, it's important to document the research conducted on the dataset; but how detailed should this be? It would be nice to be able to re-use data in another project, for slightly different purposes, ... in an easy way
- Arguments that can be used for discouraging users from withdrawing their data: scientific integrity, reproducibility of experiments
- We discussed using private contracts with data subjects (instead of simple consent forms) when possible in order to make it more difficult for them to withdraw
- Opposing "forces"
    - Consent vs legitimate interest
    - Personal interest vs public interest (in some cases the data subjects cannot (are not legally allowed to) withdraw their data)
- Technological measures: In order to avoid disseminating videos of faces of language signers, it's possible to use avatars; however
    - we are not sure the technology is mature enough
    - it is time and power-consuming
    - there's a loss of information
    - potentially the dataset cannot be re-used for other purposes
- Re-purposing the data
    - for the sake of
        - automatic sign recognition and MT
        - studying sign language variation (individual & geographic)
        - face recognition of emotions of deaf people - are the comments also applicable to other people?
    - for this, we might need in addition
        - additional biographical information
        - linking to other dictionaries

Use case 2 (DH Courses Registry)

- Clarifications:
    - user registration & tracking is only for documentation purposes; access is allowed for all
    - no enrichment of data (only what the submitter tells) but there's a checking mechanism (e.g. for URLs not working)
- Considered a rather safe dataset, similar to a database of institutional publications, given that the data are entered by users (on their own free will) and not harvested or collected by other means.
- Policy and legal framework for the APIs must be decided; think of
    - who is allowed to download
    - how are they allowed to use it?
    - is there any user monitoring?
- A page must be added with information (on allowed potential uses of the data) asking users to accept it before downloading
- Consent form of data submitters should include information on the intended uses of the API
- When a data submitter mentions the data (name etc.) of another person (course lecturer), the DH course registry manager should initiate a procedure for informing the data subject (course lecturer) for transparency reasons - but without communicating the PD of the data submitter
- The metadata should include the source of data (i.e. who submitted the course or whether it's been harvested); however the submitter's details should not be displayed on the web page of the course

- Basis for including the details of a data subject without his/her knowledge:
    - if personal information (email) is available on the internet, only need to copy; but if it is already available, what's the use of copying it? it's enough to link to the university/lecturer's page
    - if personal information is not available already over the internet, we should be reluctant to put it online
    - course material could be viewed as copyright protected work, so the name of the course lecturer-"author" HAS to be mentioned (for attribution purposes)
- Email addresses must be protected against hackers/spammers (e.g. shown as an image, with java scripts etc.)
- The DH courses registry is not like youtube etc. platforms; it is not simply a host of information; it encourages people to upload content

Use case 3 (web corpus)

- Clarifications:
    - WaCs claim to collect publicly available data (i.e. on the internet)
    - pseudonymised user profiles
    - the web corpora are not publicly downloadable; they're only available upon request on a personal basis
- Not so much a GDPR case but a huge copyright issue

Short round Case 1 (Anna Maria Bruzzone's speech archive)

- Typology: Speech archives uttered by vulnerable people in the Seventies (no consent forms available)
- Approach for legacy data: go back and ask for new consent, if possible
- The consent forms contain the right conditions (e.g. whether the interviewsd be made publicly available, under pseudonym)
- How about if only few consent forms were collected? Would it possible to invoke in that case the concept of "disproportionate effort" (GDPR, art. 14, 5b)?

Short round Case 2 (Slovenian use case)

- Consent forms for user generated content: add text for management of personal data (e.g. keeping track of user contributions, adding them to the institution's email list); users must accept the form before uploading their content
- Liabilities of the institution in case of misuse of resources: the question remains open; however, in support of science, data owner and controller should not be held responsible for eventual misuse of downloaded data from the repository

Short round Case 3 (PFC corpus)

- legacy data; difficult to go back for new consent forms, especially since a large number of countries is involved
- if there's only one distributor (Ortolang), it should have agreements from the other institutes for distributing the data
- at least those consent forms that exist, they should be documented and stored (at source institutes?)
- important notice: GDPR is not concerned with legal entities (not identifiable persons)
- how far should we go with accepting the requests for deleting/amending personal data?
- Art. 89 protects data archiving for the public interest (allows the inclusion of non-anonymized personal/ sensitive data)

Short round Case 4 (Textual input from vulnerable people)

- Not an easy case for consent forms (refugees, children, people with mental issues); try to have it in a simplified language (using even pictograms & cartoons); for refugees, important to have them in their own language, in a language they understand or use interpreters. Even in that case, one should be careful: they might be digitally illiterate.
- text or video consent form are both acceptable
- children whose legal guardians have signed the consent form can exercise the right for withdrawal when they become adult

Short round Case 5 (ReadLet project)

- Approach followed ok
- Better not to include even the number of the participant on the displayed resource; it is not needed and it has more risks of identifying the individual (if the participants' data are hacked"); having a simple "male/female" "age X" displayed is ok; on the other hand, keeping it on the displayed file allows researchers to follow the evoluation of the participant in a given span time.

Workshop participants

- Johanna Berg, Sweden (Legal expert)
- Bob Boelhouwer, DLU (LT researcher)
- Silvia Calamai, Italy (Corpus Linguist)
- George Christodoulides, Belgium (LT researcher)
- Riccardo Del Gratta, Italy (LT researcher)
- Maria Eskevich, CLARIN ERIC (LT researcher)
- Maria Gavriilidou, Greece (Corpus linguist)
- Vanessa Hannesschläger, Austria (Corpus linguist)
- Paweł Kamocki, Germany (Legal expert)
- Aleksei Kelli, Estonia (Legal expert)
- Penny Labropoulou, Greece (LT researcher)
- Laska Laskova, Bulgaria (Corpus linguist)
- Krister Lindén, Finland (LT researcher)
- Alexandros Noussias, Greece (Legal expert)
- Lene Offersgaard, Denmark (LT researcher)
- Rūta Petrauskaitė, Lithuania (Corpus linguist)
- Ineke Schuurman, Belgium (Corpus linguist)
- Serge Sharoff, UK (Corpus linguist)
- Anne-Mette Somby, Norway (Legal expert)
- Pavel Straňák, Czechia (LT researcher)
- Mateja Jemec Tomazin, Slovenia (Corpus linguist)
- Andrius Utka, Lithuania (LT researcher)
- Jurgita Vaičenonienė, Lithuania (Corpus linguist)
- Henk van den Heuvel, Netherlands (LT researcher)