SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA16105 - European Network for Combining Language Learning with Crowdsourcing Techniques
STSM title: LegiCrowd
STSM start and end date: 22/02/2020 to 08/03/2020
Grantee name: Alain Couillault

PURPOSE OF THE STSM:

Pursuing our efforts on the ethical and privacy issues regarding the collection of crowdsourced
information, the purpose was to join actors from various related fields including legal experts, linguists and mathematicians in face to face meetings. Such actors include researchers from the National University of Athens (NTUA) as well as legal partners of NTUA.

These meetings aimed at designing a more robust and computable formal description of privacy issues. This is in continuation with the previous efforts to design an ontology like description of Terms of Services by taking into account other attemps such as the P3P and the GDPR which resulted in an extended hierarchy of data (http://www.ethique-big-data.org/Data.html) and of Policies (http://www.ethique-big-data.org/PrivacyPolicy.html).

The purpose is to be able to provide a (convenient) way to annotate the complexity of Terms of Services, which can serve various purposes including legal (and notably GDPR) compliant, facilitating information exchange, preserving users' rights and, above all, enhancing transparency of Terms of Services.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

*S*everal meetings were held during this work session, between which the various actors pursued their personal explorations.
During the first phase, the team explored the relevance of existing frameworks to formally represent the content of Terms of Services.
The P3P (https://www.w3.org/P3P/) (Platform for Privacy Preferences) "*is a protocol allowing websites to declare their intended use of information they collect about web browser users. Designed to give users more control of their personal information when browsing, P3P was developed by the World Wide Web Consortium (W3C) and officially recommended on April 16, 2002.*" This approach served as a basis for the first reflections, and provided a good model for ToS representation. Unfortunately, this protocol has been retired late August 2018 due to a weak adoption by the actors. This retirement, as well as its complexity , prevented the teams to pursue this track.
Schema.org is another model at hand to represent web pages in a practical manner. The standardd is widely used due to the fact that it is supported by major search engines. Indeed it "*is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond. Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences*."
In addition, schema.org allows to extend the existing standard, and there is no existing section for Terms of Services. Collaborative work has been performed among the team to create a schema.org extension by using a dedicated platform (http://schema.legicrowd.org/). This approach was not suitable enough to annotate ToS, mainly because it lacks a performant way to describe functions ie to say, for example, that some information is stored for a certain time for a specific purpose.
Another approach consists in relying on the folksonomy induced from the crowdsourcing approach which is at the core of the ToS;DR (https://tosdr.org/) project which is embedded in so called cases (the cases are behind a login accessible page, but a copy is available here http://www.legicrowd.org/schema/schemacases.php). Discussions with the ToS;DR team have been very encouraging and we agreed to thrive to link the two approaches by relying on a linked data mechanism. The discussions among the team revealed though that, due to its bottom up approach, the set of Cases in not consistent enough but can serve as a valuable input.
The team chose to develop a different approach, which consists mainly in separating annotation and formal output (see below).

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The team developped a two stage approach for ToS annotation:
On the one hand, development of an simple annotation platform dedicated to lay men or legal experts. The platform aims at easing up the annotation process by providing a set of questions. The user/annotator is shown segments of ToSs and answers a multiple choice question. He/she can select several answers for the same questions and can or cannot answer each question. An alpha version of the platform can be found here http://www.ethique-big-data.org/legicrowddata/.

The results of the annotations are stored in a database and allows to evaluate agreement between annotators.
On the other hand, the goal is to provide different ways to export the annotations in the form of RDF triplets, which can eventually be linked to the ToS;DR cases (see above). Other formats can be considered, including CSV.
The information to be exported is to be specified, in particuliar the way the inter annotator agreement is going to be rendered in the export files.

FUTURE COLLABORATIONS (if applicable)

Future collaboration include:
- collaborative work on the design and implmentation of the platforms and of the questionnaire
- dissemination in the form of an academic paper
- submission of proposals to Call for Projects