Horizon 2020 Programme
DG CNECT
Next-generation Internet

NGI TRUST

Project acronym: **LegiCrowd Onto**

Project title: **Building an Ontology for**

**AI Based Crowdsourced annotation of Terms of Services**

Third Party: **Association des Professionnels des Industries de la Langue (APIL)**

# Final deliverable

| | |
|---|---|
| **Deliverables leader:** | **APIL** |
| **Authors:** | Alain Couillault<br>and all partners : Sofia Almpani, Theodoros Mitiskas, Alexandros Nousias, Petros Stefaneas |
| **Due date:** | February, 28th 2021 |
| **Actual submission date:** | February, 8th 2021 |
| **Dissemination level:** | confidential |

Abstract: Please provide a brief description

NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

## Disclaimer



## Copyright



This document may change without notice.

NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

## Table of contents



NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

# 1 Introduction

This is the final and sole report for the LegiCrowd Onto project. The deliverable of the project is an ontology of concepts to describe the contents of online legal documents, such as terms of services and privacy notices.

This report:

- describes what steps the team took to build the ontology
- gives a presentation of the ontology
- lists the efforts and results carried out to promote the use of the ontology towards partners

As a reminder, the three milestones of the LegiCrowd Onto project as per the initial proposal were:

- Milestone 1: a list of actors and related initiative, members of the network of actors
- Milestone 2: specification of ontology design
- Milestone 3: online wiki for ToS ontology population

# 2 Activities carried out to complete the deliverable

For sake of clarity, this section is organized in the order of milestones numbering.

### Milestone 1: network of partners

Several actions have been performed at various stages to set up a network of partners: one to one communication and presentations at conferences. We list a few of them here:

#### 1.1.1 One to one

- The partner page of the LegiCrowd website lists some of the partners http://www.legicrowd.org/index.php/partners/.
- In addition, we partnered with the French legal firm Digital&Ethics with which we submitted a project to the NGI Trust Open V2 call.
- We also made a presentation to the French Digital Ambassador team who showed great interest. Part of our development includes results of this collaboration.
- We had a discussion with the Data Protection Compliance team to envisage further collaboration.
- We discussed with the Digital Inclusion project team of the University of Louvain. Potential collaboration is twofold: set up training sessions based on the annotation platform, set up an Horizon research project when the



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

call will be published in regard to legal informatics and background technical concepts.

- We had exchanges with the schema.org community and agreed on the principle to have the resulting ontology considered as an extension of the existing de facto standard.
- We discussed with the Greek NLP community on how the LegiCrowd Onto could extend to an NLP application and joined discussions on the relevant slack channel.
- We participated in proceedings of MyData Global and explored ways of embedding LegiCrowd application in the MyData Operators architecture.

### 1.1.2 Presentations

We made several presentations at conferences:

- We presented at the Legal and AI Workshop during the 11th Hellenic Conference on Artificial Intelligence[1].
- We demonstrated the LegiCrowd Onto project to the enetcollect members on November 3rd 2020[2].
- We made a presentation at the MyData.org conference on December 10th 2020[3].
- We presented at MyData Greece 1st meet up on slack in June 2020.

## 1.2 Milestone 2: specification of ontology design

This milestone is the core of the project, which we drove in two steps.

### Step 1: define a set of tags and values

The first step consisted in defining the set of descriptors (i.e. duples {tags,values}) necessary to describe the contents of Online Legal Documents (OLDs).

An example of such duple is {collectdata, collectpersonaldatayes}. This duple means that the OLD states, regarding the collection of data, that the platform collects personal data.

In order to exhaust the set of descriptors, we iteratively created a set of {question,answer} duples, each question corresponding to a tag, and each answer corresponding to a value.

We loaded the set of questions and answers to the annotation platform made available by the APIL (see figure below) and tested the existing set by annotating existing OLDs, which lead us to add more duples.

We followed a series of alpha1... alpha2 internal tests, then internal and external Beta tests (see Gantt chart).

We called for our various contacts mentioned above to participate in the external Beta test, which gave us some very valuable feedback.

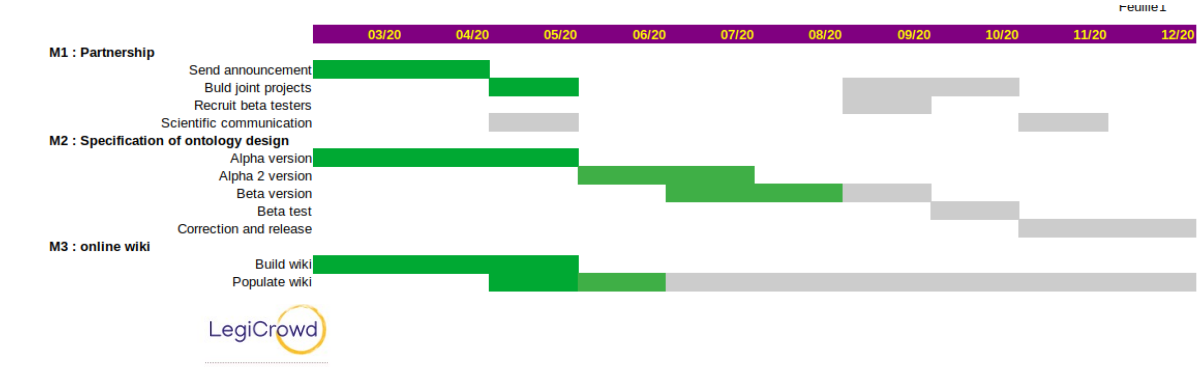
1 https://www.eetn.gr/

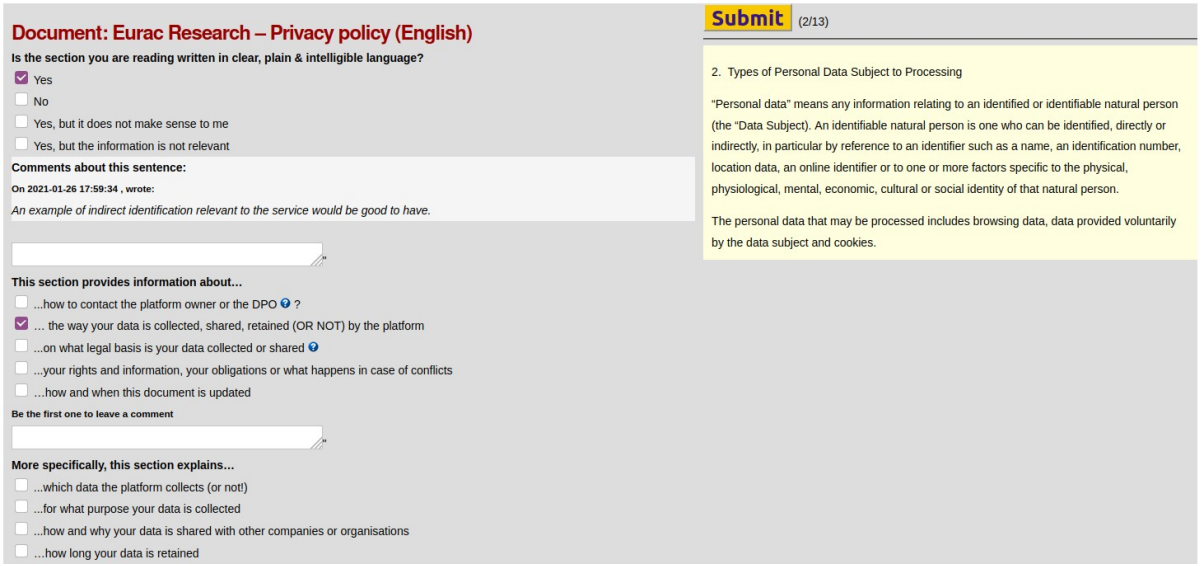
2 https://enetcollect.eurac.edu/enetcollect-related-projects/

3 https://online2020.mydata.org/

2

NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.




## Step 2: design an ontology

Once we had a large enough set of duples, we inferred an ontology to describe Online Legal Documents. We choose the schema.org approach to do so as this initiative:

- is fostered by major actors of the Internet (Google, Yandex, Bing)
- is animated by a large community
- is dedicated to the description of webpages
- uses RDF which allows to make inferences.

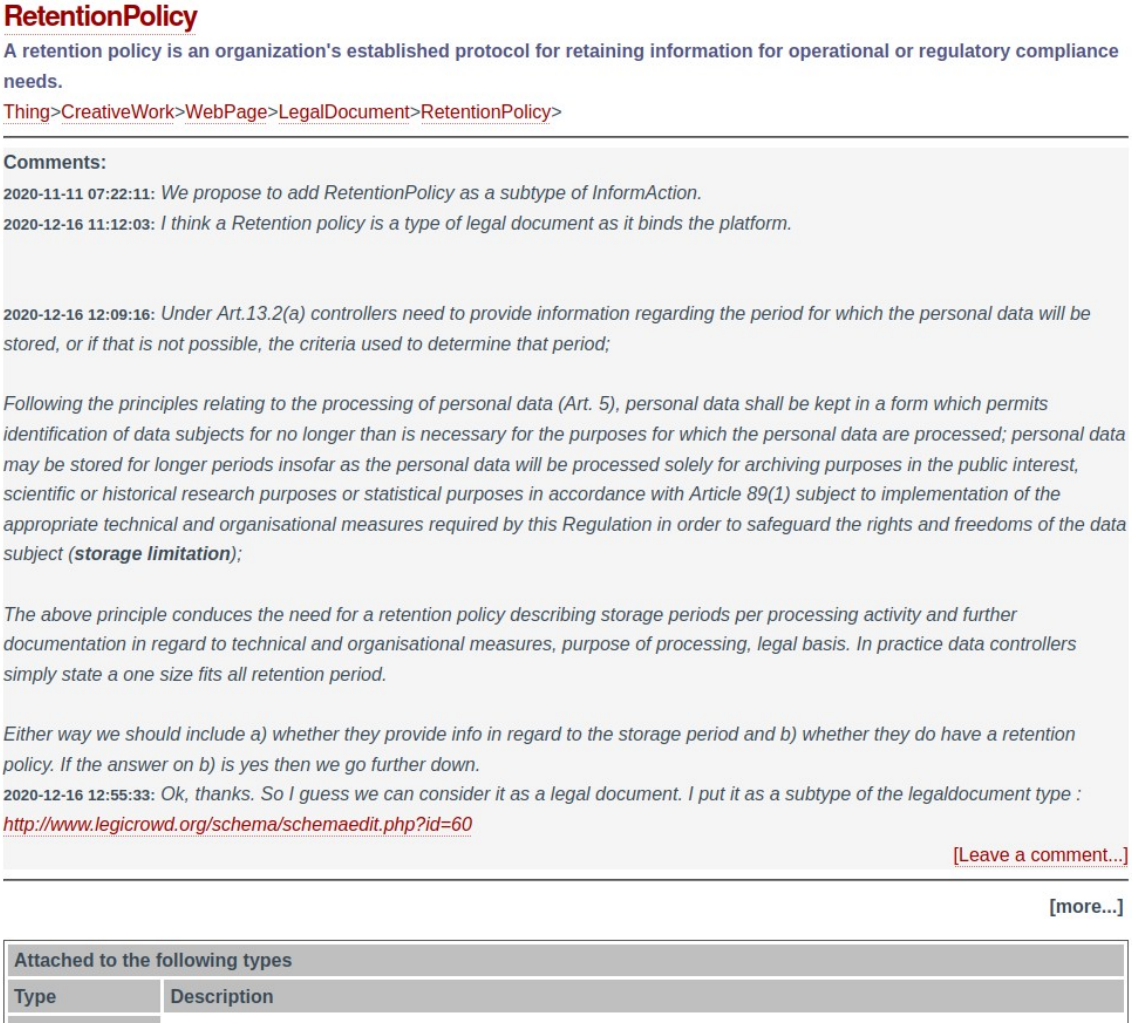
The ontology is the results of the LegiCrowd Onto project and is described in the technical description below.



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

## Milestone 3: online wiki for ToS ontology population

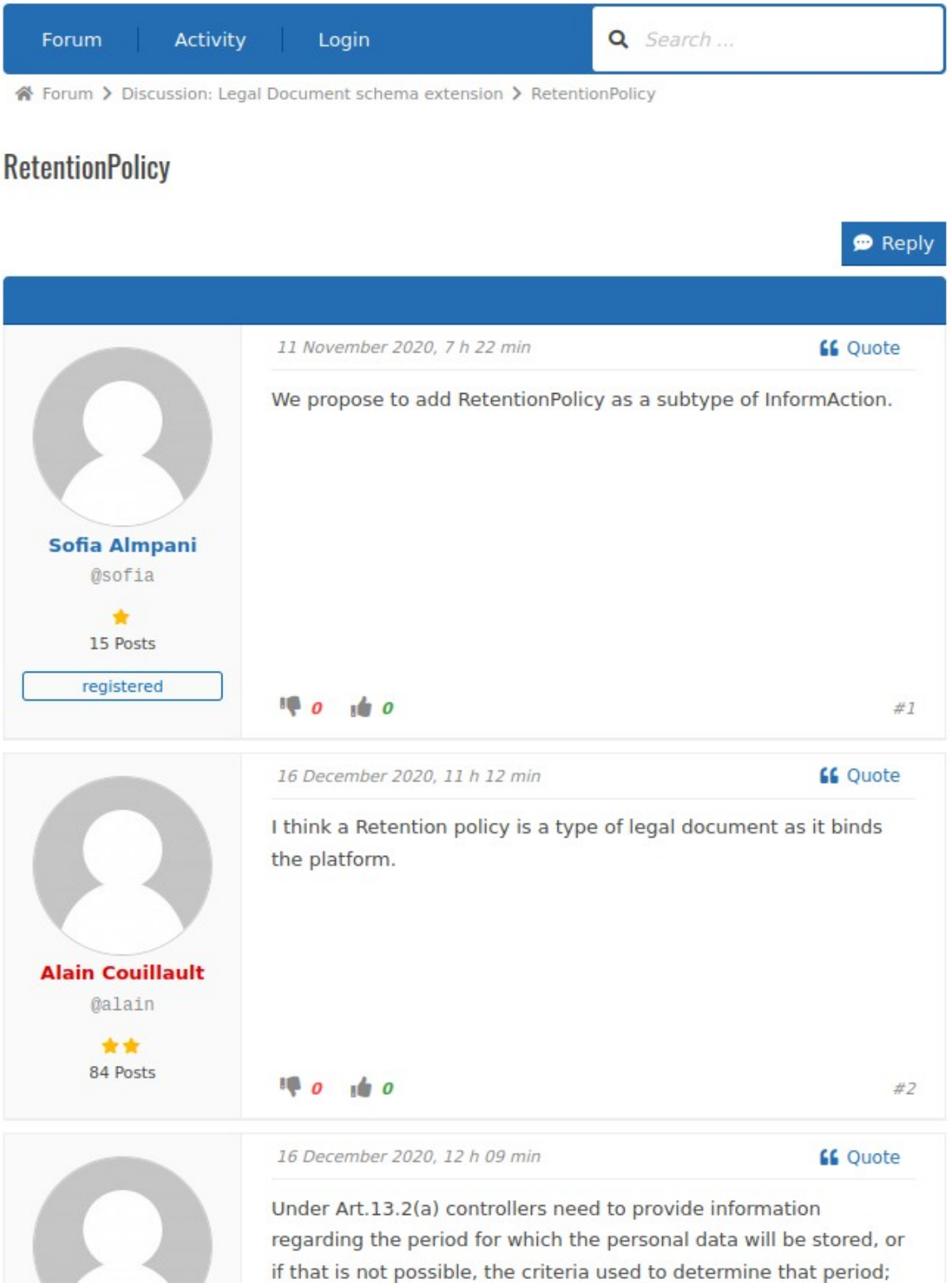
In order to exchange with the schema.org community, we set up a forum dedicated to OLD schema.org extension. Each of the changed or new type can be discussed in this forum, and the discussion is seen as comments in the description of the type. See screen capture below:




Clicking on “Leave a comment” redirects the user to the online forum:



NGI_ithubTrust project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825618.






NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

# Technical description

### Splitting Terms of Services

Since our approach was to have users annotate segments of OLDs, one of the core issues was to design and implement an algorithm to split such OLDs into consistent segments.

Several approaches have been envisaged, including machine learning or rule based natural language processing. We finally designed a program which uses the output of a program made available by the French Ambassadeur du numérique (see above) which cleans up OLDs. Our script was provided under a CC-BY-SA to their Github environment.

### Presentation of the ontology

This section provides background information for a proposed extension of schema.org dedicated to Terms of Services and Privacy Notices (i.e. Legal Documents). This extension is part of the overall LegiCrowd project and, more specifically, of the part of the project (LegiCrowd Onto) dedicated to the building of an ontology for the description of such legal documents.

This document can be found on the LegiCrowd website at www.legicrowd.org/index.php/mark-up-for-terms-of-services-and-privacy-notices/.

This extension was built by taking both a top down approach from various sources of information such as the P3P, some previous work towards to extend the P3P with GDPR specific input for data and policies, cases from ToS;DR, as well as a bottom up approach cornerstoned by the LegiCrowd annotation environment.

The LegiCrowd Onto project team is composed of Sofia Almpani, Dr. Alain Couillault (Project Leader), Theodoros Mitsikas, Alexandros Nousias, and Prof. Petros Stefaneas.

See also our partners and acknowledgements page.

## 1.1 Overview

Modelling Online Legal documents requires to handle types of objects:

- Legal documents themselves, which need to be put into the hierarchy.
- Users' data, which encompass a wide range of data types including personal data (such as a user's first name), sensitive data (for example about his beliefs or health), etc.
- Actions performed by the platform itself or the users (e.g. collect or share data, input content or cancels a contract).
- Managing users' consent to several actions performed by the platform across the consent lifecycle.

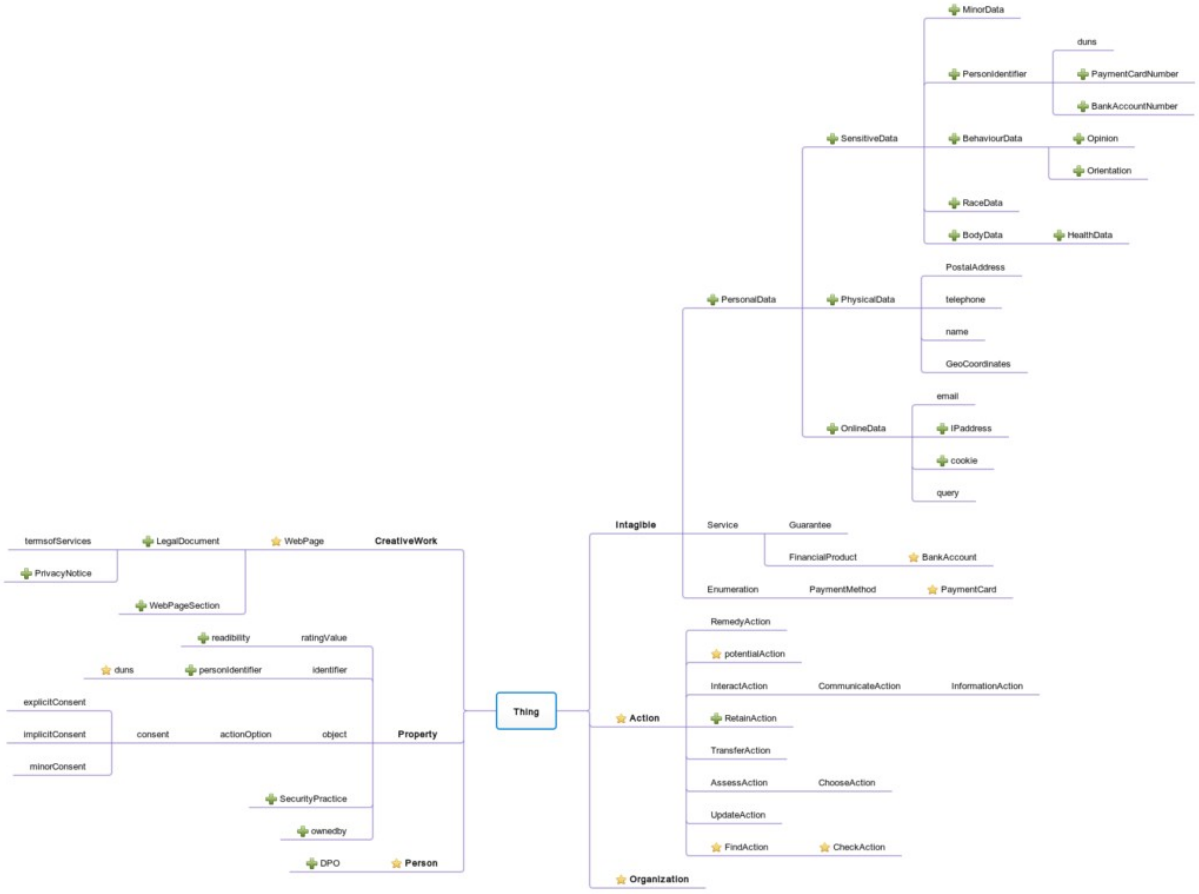


NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

- Describing the users' and the platforms' rights and commitments.

The following map gives an overview of the whole hierarchy.

- Nodes marked with a cross ( ) are types we add to the existing version of schema.org.
- Nodes marked with a star ( ) are types already in the current version of schema.org but we suggest to modify.
- All other types are standard schema.org types we use to describe legal documents.




## 1.2 Legal documents

**New types:** LegalDocument, PrivacyNotice

The current version of schema.org types Terms of Services as a Property, we propose a more generic approach: a legal document is a type of WebPage and has several subtypes, Privacy Notices and Terms of Services. The model is hence as follows (nodes with a + sign are proposed new types):




NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

In our case, PrivacyNotice and termsofServices are not different, but this model leaves rooms for specific properties for each. It has to be noted that the properties attached to the current termofServices can be inherited from the proposed LegalDocument type.

**New type:** WebpageSection

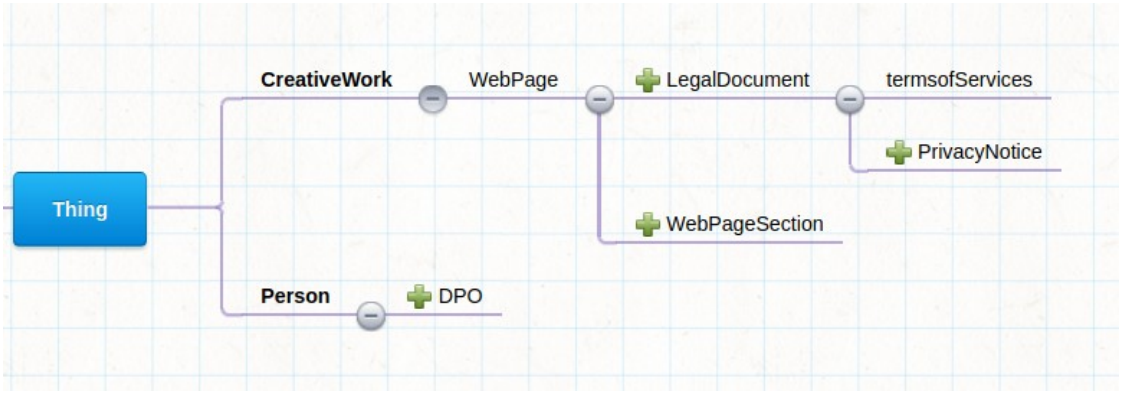
We propose to create a type WebpageSection type to describe the content of each section of a Webpage through the *hasPart* property inherited from the CreativeWork type. Though a section is not a creative work per se, the *hasPart* property provides some leeway as it is described as "Indicat[ing] an item or CreativeWork that is part of this item, or CreativeWork (in some sense)."

**New property:** *readability*, of type ratingvalue, attached to type WebPage

The existing WebPage type receives the property *readability*, which describes the readability level of the current text. It is a type of ratingvalue from which it inherits values of type text (for example: High, Low) or number (e.g. a Fleisch index).

## 1.3 Contacts

The terms of use web pages usually provide some contact information such as company email or postal address. The GDPR requires to provide contact information for the company Data Processing Officer (DPO). We hence need to add this type of contact using the *ResponsablePerson* property inherited from the CreativeWork type. We do so by creating a new type DPO as a subtype of Person.




Alternately, we could use the property *JobTitle* of the type Person but we think that this type of Responsible person is highly relevant in the context of Legal Documents. See discussion here.

The Person and Organization types of schema.org come with a large set of properties (i.e. name, email address, postal address...) which are useful for Legal Documents annotation.



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

## 1.4 Personal Data

The handling (i.e. protecting, collecting, sharing) of users' personal data is a core topic of online legal documents. A wide range of personal data can already be described with the existing types and properties, but are not marked as personal data. To achieve this, we propose to create a PersonalData type as a subtype of Intangible and to which existing or new personal data types are attached, rather than to create a specific property which appear difficult to attach to some existing data types.

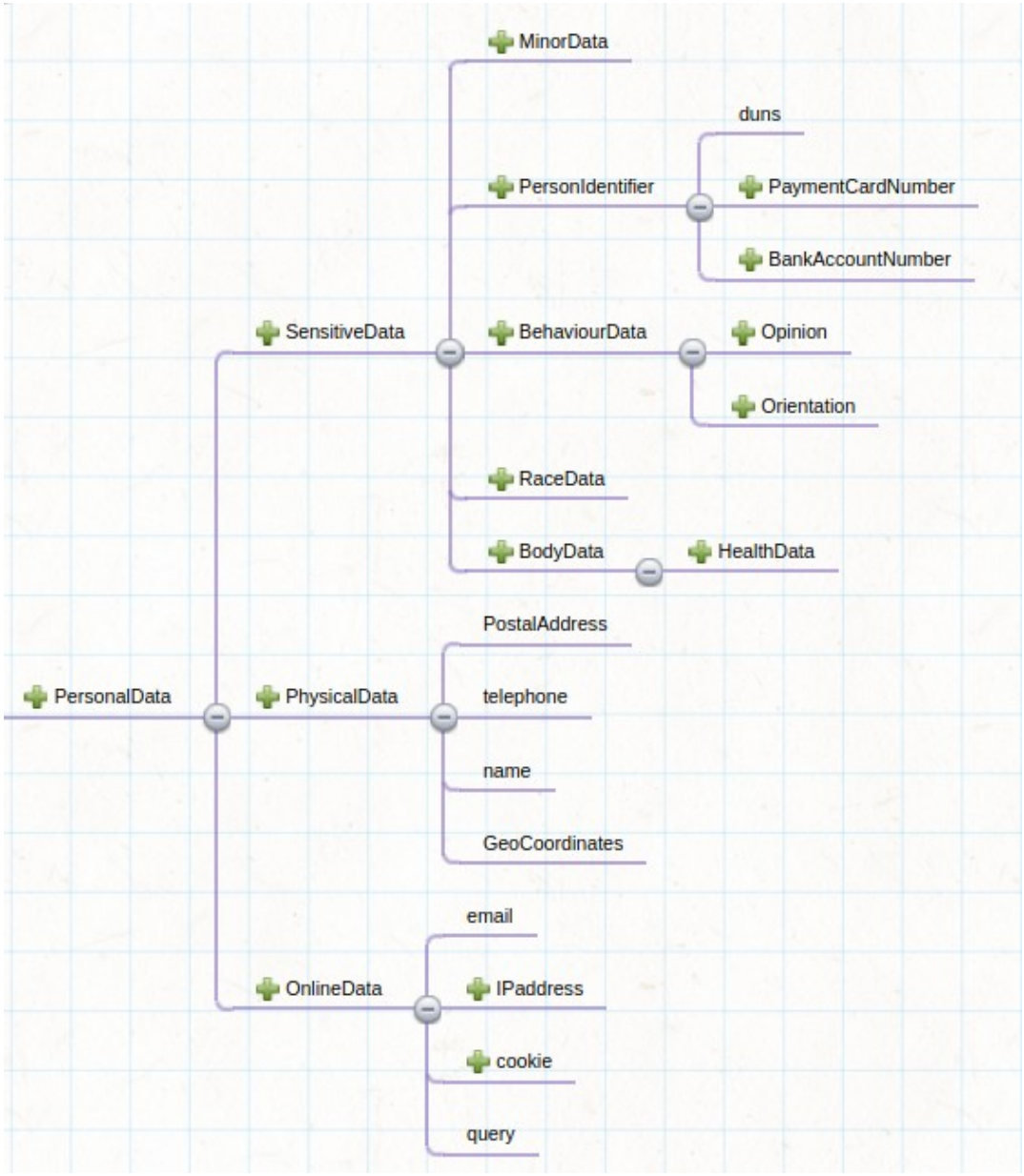
There are several types of PersonalData necessary to describe the content of Legal Document. We propose the following list of types:

- SensitiveData as described in the GDPR;
- PhysicalData related to a person in the real world;
- OnlineData related to a person's online self.

Each type of data has subtypes as described in the hierarchy below.



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.




As we see in the hierarchy above, certain types of PersonalData relates to existing schema.org types, namely:

- *query* is property defined in the current version of schema.org as "*A sub property of instrument. The query used on this action.*" and is used as a property of the SearchAction type.
- the numbers related to payment cards and bank accounts are considered as subtypes of œ&& PersonIdentifier, we then define the related BankAccountNumber and CreditCardNumber types, which can then be assigned as property of the relating existing types BankAccount and CreditCard respectively.



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

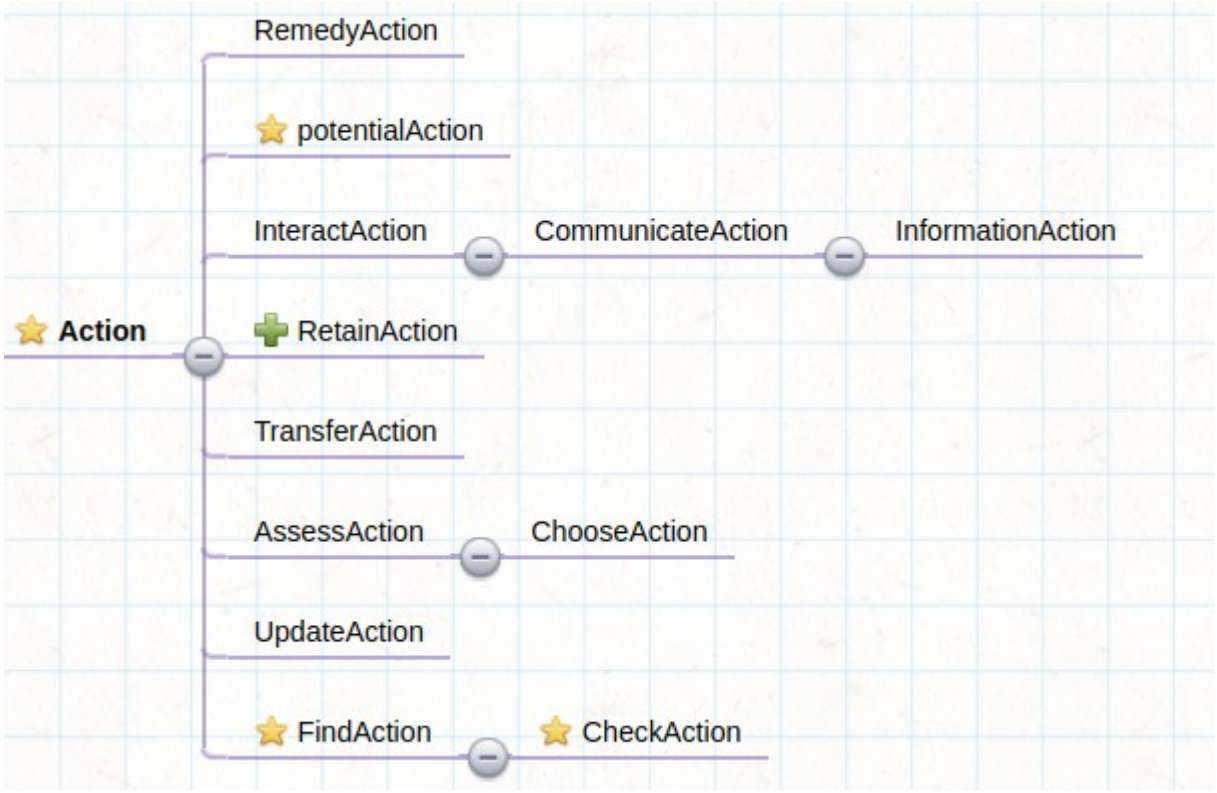
### Ownership

The notion of ownership is common to various types of information, especially, but not only, to types of personal data. The current version of schema.org does not provide such property. We hence propose to define a new property *ownedby* which can be assigned to the Thing type and hence usefully associated to any types of the hierarchy.

## 1.5 Actions on Data

We mostly rely on the large existing set of action types to describe actions necessary to tag terms of services which includes the ShareAction, potentialAction (currently first letter lower case in schema.org), UpdateAction, RemedyAction, and TransferAction types. The existing ChooseAction comes a handy to describe the user's action of giving consent.

We propose to add a RetainAction type to describe everything related to the retention of an object, and more specifically to users' data or user generated content. We also need to be able to describe the duration of the retention, and thus we propose to add the existing *duration* property to the Action type.



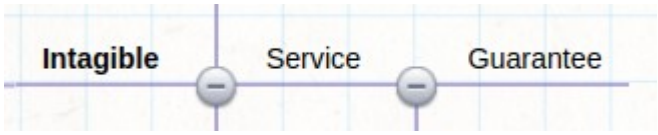

## 1.6 User's and platform's rights and commitment

We suggest adding an *hasRight* property to describe the rights attached Organization or Person, such as unsubscribe, data portability, users' data collection or sharing...



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

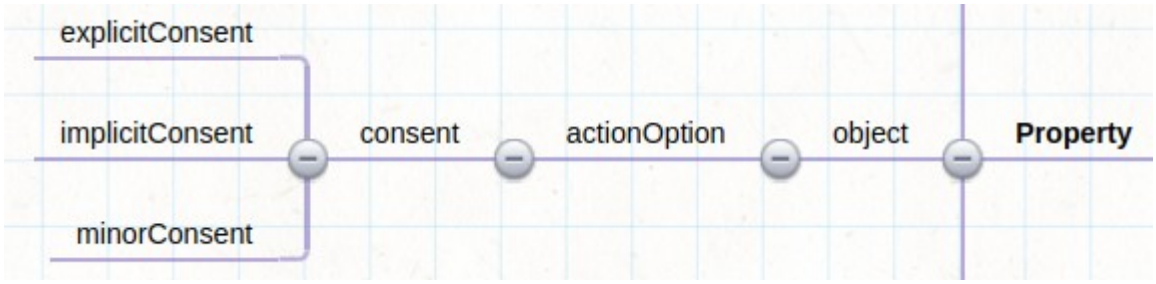
In addition, we add a type to describe the Guarantee attached to a Service to describe the platform's commitment by adding a specific type of property.




## 1.7 Managing users consents

We consider three types of consents as actionOptions:

- implicit consent,
- explicit consent,
- minor explicit consent.




This allows to describe consents to various types of actions, such as collect, share, port... data.

### 1.7.1 Checking minors consent

We need to enhance the existing FindAction type in order to cater for checking the minor's responsible person identifier. This involves both to change the description of the FindAction and CheckAction types, and to add the type Person as a property of FindAction which is then inherited to the CheckAction type.

## 1.8 Additional types of legal documents

Beyond describing terms of services and privacy notices, the LegalDocument type can be used to describe various types of documents, such as official government forms or web related documents. This could include:

- a court document to file or answer a lawsuit (1)
- a USCIS form to fill out to apply for a visa (1)
- an official CDC declaration form to fill out and give to your landlord to protect yourself from eviction (1)
- Trackers Policy (2)
- Parent Organization Terms (2)
- Parent Organization Privacy Policy (2)
- Developer Terms (2)



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

- Community Guidelines (2)
- Acceptable Use Policy (2)
- Restricted Use Policy (2)
- Commercial Terms (2)
- Copyright Claims Policy (2)
- Law Enforcement Guidelines (2)
- Human Rights Policy (2)
- In-App Purchases Policy (2)
- Review Guidelines (2)
- Brand Guidelines (2)
- Quality Guidelines (2)
- Data Controller Agreement (2)
- Data Processor Agreement (2)
- User Consent Policy (2)
- Closed Captioning Policy (2)
- Seller Warranty (2)
- Single Sign-On Policy (2)
- Vulnerability Disclosure Policy (2)
- Live Policy (2)

*(1) from* https://lists.w3.org/Archives/Public/public-schemaorg/2020Mar/0022.html

*(2) from* https://github.com/ambanum/CGUs/blob/master/src/app/types.json#L32



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

## 1.9 The whole hierarchy in the schema.org format



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

**legicrowd.org**

(Hello alain) Documentation LegiCrowd Forum TOS Annotator Logout

The whole hierarchy

[Toggle view...]

- *Thing* (changed)
  - CreativeWork
    - *WebPage* (changed)
      - LegalDocument (added)
        - termsofServices
        - PrivacyNotice (added)
        - RetentionPolicy (added)
      - WebPageSection (added)
    - WebContent
      - userGeneratedContent (added)
    - ArchiveComponent
      - ArchivesOfChangesInTerms (added)
    - Legislation
    - Article
      - SocialMediaPosting
  - *Intangible* (changed)
    - Property
      - ratingValue (added)
        - readability (added)
      - identifier
        - personIdentifier (added)
          - duns
          - BankAccountNumber (added)
          - *PaymentCardNumber* (changed)
      - instrument
        - query
      - permissionType
      - brand
      - hasRight (added)





NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

- jurisdiction
- paymentAccepted
- contactPoint
- legalName
- *industry* *(changed)*
- conditionsOfAccess
- license
- object
  - actionOption
- ownedby (added)
- SecurityPractice (added)

PersonalData (added)

- SensitiveData (added)
  - personIdentifier (added)
    - duns
    - BankAccountNumber (added)
    - *PaymentCardNumber* *(changed)*
  - MinorData
  - BodyData
    - HealthData (added)
  - RaceData (added)
  - BehaviourData (added)
    - OpinionData (added)
    - OrientationData (added)
- PhysicalData (added)
  - PostalAddress
  - telephone
  - name
  - *GeoCoordinates* *(changed)*
- OnlineData (added)
  - email
  - IPAddress (added)
  - Cookie (added)
  - query

Enumeration

- PaymentMethod
  - PaymentCard

*Service* *(changed)*

- FinancialProduct



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

- *BankAccount* (changed)
- Guarantee (added)
- Consent (added)
- ImplicitConsent (added)
- ExplicitConsent (added)
- MinorConsent (added)
- Quantity
- Duration
- *Person* (changed)
- DPO (added)
- *Action* (changed)
- RemedyAction (added)
- *potentialAction* (changed)
- *FindAction* (changed)
- *CheckAction* (changed)
- InteractAction
- CommunicateAction
- InformAction
- UpdateAction
- RetainAction (added)
- AssessAction
- ChooseAction
- *Organization* (changed)



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.

# 3 Conclusions and next steps

Because of the pandemic, we had to restrain some of our travels at a moment where it was most needed, we hence had to shift some funding from travel to people expenses and extend the project by a couple of months.

### Collaboration

During our exchanges with potential partners, we noticed a strong interest of actions related to privacy and terms of services. Be it in the academic sector (we envisage future collaborations with research teams under a soon to come Horizon umbrella) or the private sector (we already have submitted a NGI Trust with a French private company, existing Mydata.org data operators are potential users). This motivates us to pursue our efforts and set up new projects with new partners.

### Crowdsourcing

While discussing with partners, and from our own experience during the project, we found out that animating a crowd of non-expert to make annotations can be a tedious and expensive tasks. We envisage alternatives such as dedicating human resources or organizing dedicated workshops. Automatic annotation (beit machine learning or rule based) requires to have a large set of annotating OLDs and might be considered eventually.

### Legal environment

The legal environment at worldwide (i.e. Californian Privacy Rights Act) and European levels (i.e. the Data Governance Act) shows that new regulations provide opportunities for such platforms.

### Future endeavours

Our future endeavours are threefold:

- An affectio societatis has emerged between the partners of the project thanks to more than a year of working together on and off line. The team gathers all the skills to bring a product to market (management, product management, legal, marketing, Research and Development). We have submitted a DAPSI project that we see as a great potential leverage for that purpose.
- Continue the contacts with the academic world to pursue our R&D efforts.
- Ensure the developed schema.org extension be part of the schema.org de facto standard.



NGI_ithubTrust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618.