

FICHE BILAN

PROJETS AMORÇAGE 2022

Annexes:

- Programmes des manifestations scientifiques organisées dans le cadre du projet
- Offre de stage

MétaDroit

MÉTAdonnées et ontologies dans le champ du DROIT

1 – Bilan des actions réalisées

Dans le cadre du projet MétaDroit, nous avons organisé deux séminaires.

Le premier séminaire s'est déroulé à la faculté de droit de Nantes Université le 18 mars 2022 en trois temps : présentation du projet, contributions scientifiques, discussions entre les membres du projet et les invités sur les intérêts de chacun (cf. programme en PJ). Les aspirations universalistes des métadonnées, leur délicate évolution dans le temps au sein du champ mouvant qu'est le droit, le manque de transparence sur leurs usages ou encore leur emploi à des fins de traçage ont été autant d'enjeux abordés. Réunissant sur place, et à distance, une vingtaine de participants provenant de champs disciplinaires différents (droit, informatique, langue et économie), ce premier séminaire de travail nous a permis de délimiter notre objet d'étude, d'envisager la place de l'interdisciplinarité dans ce cadre et de poser les jalons des travaux suivants de MétaDroit. En effet, le séminaire a été l'occasion de dégager un premier axe prometteur de travail : celui lié à l'automatisation, c'est-à-dire de l'utilisation de systèmes informatiques simples ou complexes afin d'effectuer certaines tâches en lien avec les métadonnées et ontologiques ou permises par ces dernières. Cette automatisation concerne différents aspects : la création d'ontologie, l'annotation des textes juridiques par les métadonnées ou encore les usages automatisés permis grâce aux métadonnées et ontologies à l'image de la traduction juridique ou de l'analyse de données en grand nombre respectant le droit de la protection des données personnelles. La question de l'automatisation a donc été retenue comme piste de travail, d'inspiration pour l'après « MétaDroit ».

Le second séminaire du projet s'est déroulé le 8 novembre 2022 au LS2N. L'objectif était de croiser les points de vue sur l'usage de métadonnées avec des travaux étrangers, nationaux, en informatique et en droit. Nous avons eu la chance d'écouter Sabrina Kirrane sur la thématique suivante ; "*The linked legal data landscape: linking legal data across different countries*"; Mounira Harzallah sur les "Ontologies : Intérêt des ontologies et méthodes de leur apprentissage semi-automatique". Enfin nous avons reçu Mélanie Dulong de Rosnay « Legal ontologies, From copyright to data commons » (cf. programme en PJ). Réunissant sur place et à distance une quarantaine de personne (de champs disciplinaires différents), ce second séminaire a été l'occasion de consolider le réseau établi et, sur le fond, d'interroger la logique d'automatisation, les politiques d'annotations officielles en droit au sein de l'Union européenne ou encore les moyens d'investiguer ces questions en droit.

Autres résultats

Captations :

Les séminaires ont fait l'objet de captations. Elles n'ont pas encore été diffusées. Elles serviront à tout le moins comme support pédagogique de référence pour les étudiants (M2 et Doctorants) travaillant dans le champ des métadonnées et désireux de saisir les enjeux techniques ou les enjeux juridiques s'y rattachant.

Autre intervention s'appuyant sur les travaux de Métadroit :

BERNELIN M. « **Les métadonnées, des données comme les autres ?** », Colloque interdisciplinaire *A l'épreuve des données : Sensibilité, interprétation, appropriation des données numériques*, 27-28 octobre 2022, Université Catholique de Lille.

2 – Bilan scientifique

L'objectif du projet Métadroit était de créer un réseau d'intérêt autour de la thématique des métadonnées juridiques, thématique très peu analysée en droit. Ce pari a fonctionné au-delà de nos attentes, d'autres disciplines que le droit et l'informatique ayant été représentées lors de nos échanges. Nous avons également pu nouer des liens à l'étranger.

Outre les éléments présentés dans la partie 1, laquelle dresse un bilan scientifique par séminaire organisé, notre projet et nos échanges ont donné lieu à la création d'un nouveau partenariat de recherche. En effet, sous l'impulsion de Madame Mounira Harzallah, Métadroit se prolonge par une approche pratique visant à constituer de manière automatisée un corpus de métadonnées juridiques pour décrire le droit relatif à l'Intelligence artificielle. Intitulé « **Combining machine learning techniques for "Regulation of Artificial Intelligence" ontology building** », cette recherche a débuté en décembre 2022 et se prolonge en 2023 avec le recrutement d'un étudiant de Master en informatique pour comparer et combiner les méthodes afin de constituer de manière semi-automatisée un corpus de métadonnées. Les travaux impliquent une réflexion interdisciplinaire sur les concepts principaux devant figurer dans les métadonnées, leurs rapports les uns aux autres ainsi qu'une évaluation de leur extraction automatique au sein de la Proposition de règlement de l'Union européenne sur l'Intelligence artificielle. Ces travaux, une fois terminés, seront concrétisés notamment par une publication. L'offre de stage relative à ce projet est annexée à ce bilan.

3 – Perspectives

Les perspectives à l'issue de notre projet Amorçage Métadroit, outre les travaux sur la constitution d'un corpus de métadonnées juridiques, visent à prolonger le projet par le dépôt de candidatures à des APP ciblés : **ANR JCJC** et **ERC Starting Grant** (dépôts par Margo BERNELIN).

MétaDroit

Premier séminaire

MÉTAdonnées et ontologies dans le champ du DROIT

```
rdfs:
```

```
tp://cali.priloo.univ-
```

```
tp://cali.priloo.univ-nantes
```

```
tp://api.kasabi.com/dataset/nasa/apis/
```

```
s:label "NASA Space Flight & Astronaut
```

```
s:comment "Conversion of various NASA data
```

```
raft data from the NSSDC master catalog." ;
```

```
l:hasPolicy <http://cali.priloo.univ-
```

```
fr/api/ld/licenses/65927752496731336041331568383952761331>
```

```
tp://cali.priloo.univ-nantes.fr/api/ld/licenses/MIT>
```

```
s:label "MIT" ;
```

```
od:licensingTerms <https://opensource
```

```
l:Duty cc:Notice ;
```

```
l:Permission cc:CommericalU
```

```
cc:DerivativeWorks,
```

```
cc:Distribution
```

```
rd cc:Reproduced "MIT" ;
```

```
odrl:
```

```
14iod.licensingT
```

```
odrl:Duty c
```

```
odrl:Pe
```

FACULTÉ DE DROIT

Salle 220

Nantes Université

Disponible en visioconférence

Contact et inscription : margo.bernelin@univ-nantes.fr

Programme :

14h-14h30 : Accueil et présentation du Programme MétaDroit

14h30-15h15 : Présentation scientifique par Patricia Serrano Alvarado (MCF HDR, LS2N, Nantes Université), « Sur les ontologies pour la protection de données »

15h15-15h30 : pause

15h30-16h15 : Présentation scientifique par Margo Bernelin (CR CNRS, DCS, Nantes Université), « Les enjeux juridiques du recours aux métadonnées et ontologies dans le champ du droit »

16h15-17h : Tour de table avec les participants : intérêts de recherche, opportunités de réseaux, interventions scientifiques à venir.

Seminar n°2



AAP «Amorçage»

MétaDroit

Metadata and ontologies in the field of law



13H30-16H30
MARDI 8 NOVEMBRE
2022

LS2N,
FACULTÉ DES SCIENCES ET TECHNIQUES

Amphithéâtre du bâtiment 34

2, Rue de la Houssinière, 44322 Nantes, France

Disponible en visioconférence

Contact et inscription : margo.bernelin@univ-nantes.fr

INTERVENANTES :

- Sabrina Kirrane, Institute for Information Systems & New Media, Vienna University of Economics and Business
- Mourina Harzallah, Laboratoire des Sciences du Numérique de Nantes, Nantes Université
- Mélanie Dulong de Rosnay, Centre Internet et Société (CNRS)

Avec le soutien de l'AAP «Amorçage» de la MSH Ange-Guépin
Dans le cadre du projet MétaDroit - MÉTAdonnées et ontologies dans le
champ du DROIT



Master Internship

Title: Combining machine learning techniques for “Regulation of Artificial Intelligence” domain” ontology building

Research laboratories: DUKe/GDD, LS2N (www.ls2n.fr) & “Law and Social Changes” laboratory(dcs.univ-nantes.fr)

Supervisors:

Margo Bernelin (Margo.Bernelin@univ-nantes.fr), “Law and Social Changes” laboratory

Mounira Harzallah (mounira.harzallah@univ-nantes.fr), DUKe, LS2N

Patricia Serrano Alvarado (Patricia.Serrano-Alvarado@univ-nantes.fr), GDD, LS2N

Context

The master internship will be carried out in the context of the MétaDroit project (metadata and ontologies in the field of law) within a collaboration between the team InnovSanté of the “Law and Social Changes” laboratory (dcs.univ-nantes.fr/) and the teams DUKe and GDD of LS2N.

In the project, we are interested in ontology building from texts for the “Regulation of Artificial Intelligence” domain. Indeed, the regulation of Artificial Intelligence is emerging in Law prompting legal discussions. Currently, there isn’t an ontology that formalises the knowledge of the domain and allows us for example to identify high-risk AI systems and infer for each one a set rules/constraints that should be respected or a set of horizontal obligations on its providers.

Recently, the European Commission has elaborated three texts (i.e. corpus) that aim at regulating all dimensions of AI: from conception, to marketing and liability (eur-lex.europa.eu/legal-content/FR/ALL/?uri=CELEX%3A52021PC0206). We seek to process the corpus and test on it several machine learning approaches for ontology learning and then to compare the performance of these approaches in order to propose a combined one for building an ontology for the “Regulation of AI”.

Problematic

The popularity of ontologies and the easy access to a large number of textual resources have strongly motivated the automatic construction of ontologies using artificial intelligence techniques. Two types of construction approaches are distinguished: pattern-based approaches and distributional approaches and [Xu et al., 2019, Chen et al. 2020]. Pattern-based approaches show quite high precision whereas their recall is low because of the large variability in natural language for expressing a meaning. Distributional approaches can be supervised or unsupervised. Supervised approaches perform well, however they are rather sensitive to the distribution of training dataset, making question able their reliability. In addition, training dataset building is time-consuming. Unsupervised approaches and more specifically clustering based approaches don’t require a training dataset and allow to consider a large amount of data. However, they face two main difficulties: the cluster labeling and the formation of semantically consistent clusters relevant to the ontology domain. Though these approaches seem to be complementary, there has been rather little work on integrating them.

Master internship purpose

The purpose of the master internship is to compare and combine four approaches for the semi-automatic construction of an ontology for the “Regulation of AI” domain. The work will be focused on the extraction of two ontology components: concept and hypernym relationship. Four approaches will be compared: Core concept seeded-LDA [Huang et al.

2021]¹, similarity measure based approach [Albukhitan, 2017], word-embedding-based-supervised approach[Mikolov et al. 2013]² ³, and hypernym-pattern-based approach [Alaa Aldine et al. 2021]. The combined approaches will be a core ontology driven approach and its results will be enriched by knowledge from knowledge graphs, such as DBpedia.

Work outline

1. State of the art : core-concept seeded LDA, pattern-based approach, Combined approach.
2. Building a benchmark (a gold standard ontology) from the “Regulation of AI” corpus.
3. Testing four approaches for ontology learning and evaluation of their results.
4. Proposition of a combined approach for learning an ontology for the domain of “Regulation of AI”.
5. Implementation and validation of the combined approach on the “Regulation of AI” corpus.

Candidate Profile : Knowledge on Data mining/Machine Learning, Knowledge on semantic web and NLP will be strongly appreciated but not mandatory, Knowledge in programming languages mainly Python.

Expected starting date: around February 2023.

Gratification: 6 months ~ 550 € / month

To apply send your CV, a cover letter and your transcript of records of the tree last years to margo.bernelin@univ-nantes.fr, mounira.harzallah@univ-nantes.fr and patricia.serrano-alvarado@univ-nantes.fr

References

[Huang et al, 2021] H. Huang, M. Harzallah, F. Guillet, Z. Xu, Core-Concept-seeded LDA for Ontology Learning. KES2021 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2021.

[Albukhitan, 2017] Albukhitan, S., H.T.A.A., 2017. Arabic ontology learning using deep learning, in: Proceedings of the International Conference on Web Intelligence, pp. 1138–1142.

[Alaa Aldine et al. 2021] A.Alaa Aldine, M. Harzallah, G. Berio, N. Béchet, A. Faour. A 3-phase approach based on sequential mining and dependency parsing for enhancing hypernym patterns performance. In Knowledge Engineering Review; Cambridge University Press (CUP), vol. 36, 2021.

[Chen et al. 2020] Chen, C., Lin, K., Klein, D.: Inducing Taxonomic Knowledge from Pretrained Transformers. arXiv preprint arXiv:2010.12813, 2020.

[Seitner et al., 2016] Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., and Ponzetto, S. P. (2016). A large database of hypernymy relations extracted from the web. In LREC.

[Xu et al., 2019] Xu, Z., M. Harzallah, F. Guillet, and R. Ichise. Modular Ontology Learning with Topic Modelling over Core Ontology. Procedia Computer Science 159 (2019): 562-571.

¹ <https://github.com/jason-huanghao/Core-Concept-Seeded-LDA>.

² <https://radimrehurek.com/gensim/>

³ <http://scikit-learn.org/>

[Mikolov et al. 2013] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In NIPS pp 3111-3119.

[Shwartz et al., 2016] Shwartz, V., Y. Goldberg, I. Dagan. Improving hypernymy detection with an integrated path-based and distributional method. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, 2016.