RAPPORT D'ACTIVITÉ

OLKi

Open Language and Knowledge for Citizens

INVESTIR L'AVENIR

UNIVERSITÉ DE LORRAINE

LORRAINE Université d'Excellence

IMPACT OLKi

ÉDITO

Christophe Cerisara
Chargé de recherche CNRS au Loria
Porteur scientifique du projet

Bien que la période soit particulièrement difficile et ne se prête guère aux réjouissances, le projet OLKi a accompli un chemin conséquent, grâce aux efforts de toutes et de tous depuis son démarrage en octobre 2018.
Ce chemin concerne d'abord les contributions scientifiques du projet, que vous retrouverez dans les pages qui suivent.

Mais au-delà de ces avancées, je retiens avant tout la cohésion et les échanges qui ont jalonné ces deux années. La principale force du projet OLKi est la communication entre les disciplines qui s'y est développée : je l'ai ressentie comme omniprésente et particulièrement forte au sein de notre consortium. Ces échanges font régulièrement émerger de nouvelles collaborations et se traduisent dans les faits : ainsi, 83% des thèses sont co-encadrées par deux laboratoires appartenant à des pôles scientifiques différents.

Mais l'essentiel va au-delà des indicateurs : c'est l'influence que ces échanges ont sur chacun d'entre nous, influence discrète mais grandissante, génératrice de curiosité intellectuelle et d'opportunités.
Vous découvrirez dans ces pages quelques étapes de ce chemin, mais il y a bien entendu beaucoup plus dans OLKi que ne le résume ce livret ; j'espère que le parcourir vous incitera à aller voir plus loin, en commençant par toutes nos ressources, notamment vidéos, sur le web.

Au nom de tout le comité OLKi, je remercie l'Université de Lorraine et LUE pour nous avoir accordé leur confiance et les directions des laboratoires partenaires pour leur précieux soutien; je ne remercierai jamais assez tous les membres du comité opérationnel de OLKi, en particulier la chargée du projet OLKi, ainsi que l'ensemble de la communauté scientifique qui nous accompagne.

Merci à vous !

SOMMAIRE

Le masculin est utilisé par convention dans ce document, mais désigne des hommes et des femmes.

CHIFFRES-CLES

5 laboratoires du site lorrain

Archives Henri-Poincaré Philosophie et Recherches sur les Sciences et les Technologies

atilf Analyse et traitement informatique de la langue française

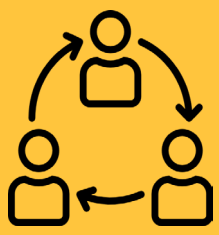
crem centre de recherche EA 3476 sur les médiations communication, langue, art, culture

Institut Élie Cartan

Loria

Budget
1 531 858 €

Plus de 100 personnes impliquées



3 rencontres chercheurs - acteurs économiques

Soutien à 2 conférences internationales

Création d'un séminaire philosophie-informatique 5 sessions en replay

>15 événements

2 écoles thématiques de portée internationale

½ journée d'étude thématique sur les corpus des réseaux sociaux

4 journées scientifiques nationales

1 plateforme

olki.loria.fr/platform

>120 productions associées

hal.archives-ouvertes.fr/IMPACT-OLKI

LE PROJET

Le contexte Lorraine Université d'Excellence

Initiée au printemps 2016, l'initiative Lorraine Université d'Excellence (LUE) a été retenue dans le cadre de l'appel d'offres du gouvernement français PIA2 IDEX/I-SITE, avec le label I-SITE (Initiatives - Science – Innovation – Territoires – Economie). Le projet Open Language and Knowledge for citizens (OLKi) est le projet du programme IMPACT de LUE répondant au défi de l'ingénierie des langues et des connaissances.

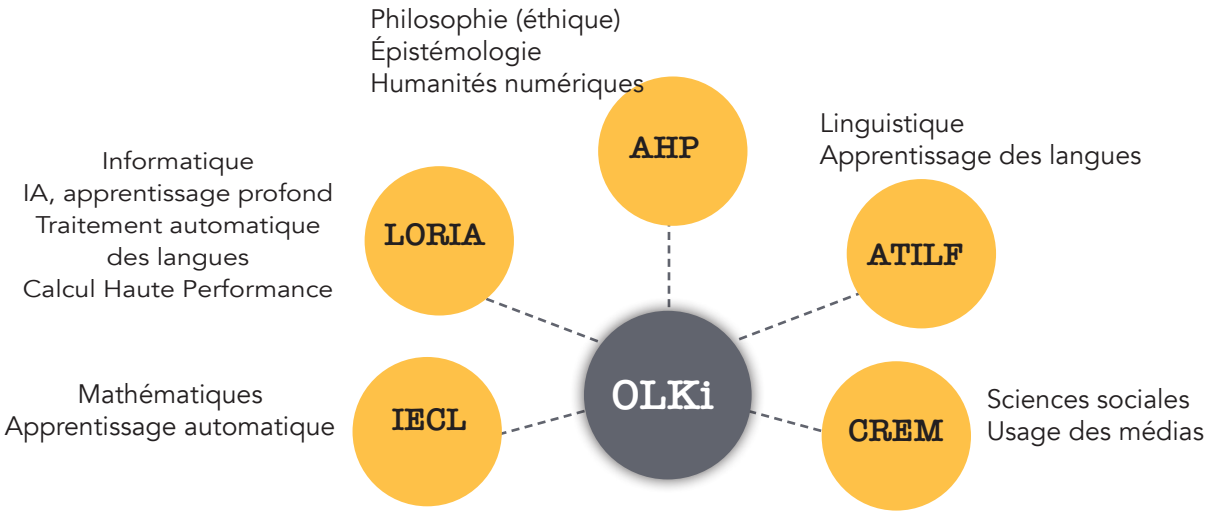
La vision
Un projet interdisciplinaire autour de l'IA et des données langagières

Les méthodes d'apprentissage profond (deep learning), omniprésentes aujourd'hui dans les domaines applicatifs de l'IA se nourrissent de grandes quantités de données. Les données acquièrent ainsi aujourd'hui une valeur stratégique pour l'IA comparable à celle du pétrole pour l'énergie. Or, ces richesses, même lorsqu'elles sont produites en Europe, sont principalement contrôlées par des acteurs étrangers et échappent aux scientifiques français et européens. Il nous faut inventer des solutions pour que d'une part les citoyens gardent le contrôle de leurs données tout en accordant d'autre part aux scientifiques la possibilité de réaliser des recherches éthiquement irréprochables.

Le projet OLKi a pour mission de concilier deux problématiques : concevoir de nouveaux algorithmes d'apprentissage automatique dédiés à l'extraction des connaissances à partir de données langagières, et proposer des solutions qui garantissent un contrôle équitable, ouvert et partagé des données ainsi qu'une utilisation de ces données qui respecte le citoyen et sa vie privée.

Démarrés en 2018, les travaux sont menés au sein du projet OLKi de manière interdisciplinaire par des spécialistes en IA, apprentissage, mathématiques, linguistique, philosophie et sciences de l'information et de la communication.

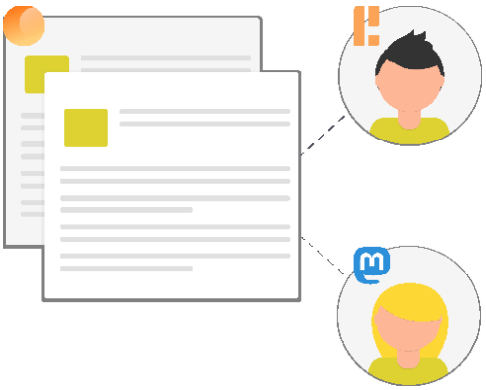
Le consortium : un projet porté par 5 laboratoires



Budget : 1 531 858 euros

LE PROJET

Organisation



Comité opérationnel

Le comité opérationnel du projet OLKi est constitué du responsable scientifique du projet, de deux représentants de chaque laboratoire du consortium, et de la chargée de projet.
Le comité se réunit une fois par mois depuis le début du projet. Des présentations scientifiques invitées ont lieu depuis 2020 afin de s'acculturer aux différentes disciplines et travaux effectués dans le cadre du projet.
Sont également invités et tenus informés, les directions des laboratoires, les directions des pôles scientifiques Automatique, Mathématiques, Informatique et leurs interactions (AM2I) et Connaissance, Langage, Communication, Sociétés (CLCS), un représentant de la Direction de l'Entrepreneuriat et des Partenariats Socio-Economiques (DEPAS), un représentant de la Direction de la Recherche et de la Valorisation (DRV), et une représentante de la Satt Sayens.

Organisation

Le projet est structuré autour d'un axe relatif à la gestion de projet et au développement d'un outil, de 3 axes scientifiques (work packages – WP ; tâches – T), et de 2 axes transversaux.

WP0 Gestion de projet & Plateforme fédérée

- **T0.1** Gestion de projet
- **T0.2** Plateforme fédérée

✓ Ouvrir et faciliter la communication scientifique autour de ressources langagières
✓ Partager des ressources scientifiques (corpus, articles, vidéos, outils d'analyse, …)
✓ Construire à partir du Fediverse, un réseau international et citoyen

WP1 Connaissance et ingénierie

- **T 1.1** Réflexions éthiques et épistémologiques sur les données, l'information et les connaissances
- **T 1.2** Extraction automatique des connaissances
- **T 1.3** Transparence des algorithmes

WP2 Langage

- **T 2.1** Production de ressources linguistiques
- **T 2.2** Traitement automatique du langage naturel
- **T 2.3** Apprentissage des langues

WP3 Défis sociétaux

- **T 3.1** Production de ressources ouvertes pour les citoyens
- **T 3.2** Pratiques d'information et acquisition de connaissances
- **T 3.3** Epistémologie

Axes transversaux: Ethique & Apprentissage automatique

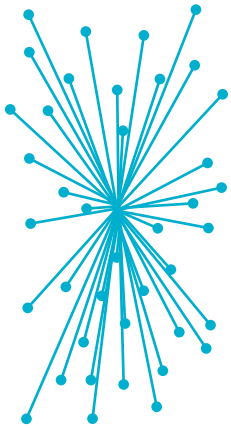
olki.loria.fr/platform

LA PLATEFORME

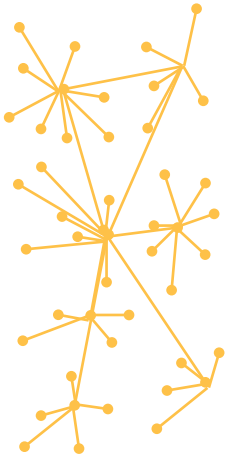
Une plateforme alternative intégrée à un mouvement citoyen de grande ampleur

La plateforme développée par le projet OLKi a pour finalité de créer un écosystème ouvert et fédéré favorisant le travail et les échanges scientifiques, notamment autour de ressources langagières. Elle permet d'héberger et diffuser des ressources scientifiques liées au langage et aux connaissances qui en sont extraites. Elle s'interconnecte aux nœuds du Fediverse (ensemble de serveurs interconnectés formant un réseau social ; https://en.wikipedia.org/wiki/Fediverse) et ajoute aux ressources qui y existent déjà (musique, blogs, vidéos…) une dimension recherche et connaissances scientifiques.

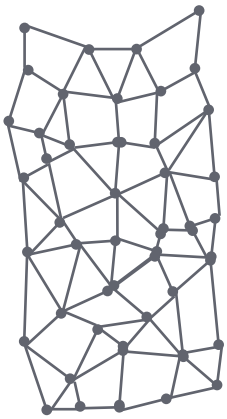
Au-delà des progrès en termes de contrôle, d'éthique, d'ouverture, de transparence et de respect de la vie privée, la plateforme résoudra en partie des problèmes de nombreuses plateformes scientifiques actuelles, dont la maintenance à long terme, le passage à l'échelle, la réduction des coûts, le contrôle des fournisseurs de données et l'interaction entre recherche et citoyens.
Le développeur, assisté d'un groupe de travail, a publié une première version utilisable de la plateforme en mai 2020 après 15 mois d'effort de développement. Elle implémente les deux fonctionnalités fondamentales: 1) la fédération des listes de corpus et 2) la fédération des commentaires. Grâce à ces deux fonctionnalités, la diffusion de corpus en cours de constitution et gérés localement devient possible via un réseau décentralisé, ainsi que les interactions de la communauté autour d'un corpus donné via les réseaux sociaux libres.



Réseau centralisé



Réseau décentralisé (ou fédéré), le Fediverse



Réseau distribué (ou pair-à-pair)

Représentations d'architectures de réseau

Profil d'utilisateur potentiel

ARCHETYPE : **Chercheur en linguistique**

AXE DE RECHERCHE : **Etude du lexique de la langue général et de la langue de spécialité**

CORPUS

DONNEES : **Textes (journaux, revues, données issues du web)**

METADONNEES : **Année de production ou récupération, langue, auteur natif ou non, sinon dépend fortement de la question de recherche**

DECOUVERTE : **Sur ORTOLANG et Redac**

CONSTITUTION : **Récupération, puis conversion en XML, puis annotation, segmentation**

DIFFUSION : **Sur ORTOLANG**

COLLABORATION

COLLABORATEURS : **1-10 personnes**

CORPUS PUBLICS : **Oui pour la plupart, certains très spécifiques ne sont pas accessibles à tout le monde**

AUTRES ECHANGES : **Avec des scientifiques pour les former à la constitution et l'exploitation de corpus, avec des étudiants**

OUTILS

- **TXM** : pour l'annotation
- **Concordanciers** : pour explorer un corpus
- **Allegro** : pour la récupération d'exemples, fonctionne avec un workflow
- **Mails et oral** : pour les discussions
- **Git** : pour échanger des fichiers et les discussions via les tickets

BESOINS

- un outil de démonstration, pour montrer le déroulement et le résultat des traitements sur des corpus
- ordonner les documents d'un corpus, y ajouter des métadonnées et liens vers des articles ou data papers
- effectuer une recherche dans les discussions
- des formations, sessions pratiques, tutos sur des outils d'exploitation de corpus
- importer des métadonnées, un corpus, une sous-partie d'un corpus

PLATEFORME OLKi

GARANTIE POUR L'UTILISATION D'UNE PLATEFORME : **Protection des données des utilisateurs et des données de recherche, libre et gratuite pour tous**

POUR LA DECOUVERTE DE CORPUS :

POUR LA DISCUSSION AUTOUR DE CORPUS :

POUR LA DIFFUSION DE CORPUS :

LA RECHERCHE

WP1 - CONNAISSANCE ET INGÉNIERIE

Mots-clés : extraction automatique du savoir, information et data, transparence des algorithmes

Le WP1 s'intéresse à la question de l'apprentissage de représentation intégrant une connaissance expert pour des données de nature très diverses : données textuelles, traces numériques en e-éducation, graphes linguistiques, réseaux sociaux...

Le point commun à tous les travaux menés au sein de ce work package est une modélisation interprétable de la structure fine des données. Les travaux menés au sein de la communauté OLKi ont permis d'apporter des contributions très diverses : modélisation de la structure discursive de textes via des approches symboliques entre linguistique et fouille de données, traitement de données de e-éducation avec comme objectif d'étudier le lien entre des descriptions d'étudiants et leurs parcours, recherche d'information dans la littérature scientifique via des plongements de graphes hétérogènes entre linguistique et apprentissage statistique, modélisation de la persistance de l'information dans les réseaux sociaux via la notion de longue mémoire.

Thèse Fouille de textes au niveau discursif (2018-2021) - Laurine Huber

Sous la direction de Yannick Toussaint (LORIA- équipe Orpailleur) et Mathilde Dargnat (ATILF - équipe Discours)

L'objectif de cette thèse est de montrer dans quelle mesure la structure des textes peut être utilisée pour améliorer les tâches de fouille de texte en combinant des approches symboliques et numériques. Pendant la première moitié de cette thèse, je me suis concentrée sur les structures du discours et de l'argumentation. J'ai développé des outils et des méthodes pour étudier si les structures construites à partir de deux formalismes distincts (du discours et de l'argumentation) sont liées, et de quelle manière. J'ai proposé deux approches qui tirent parti des techniques de fouille de données pour découvrir des alignements de sous-graphes du discours et de l'argumentation à partir d'un corpus annoté selon les deux formalismes. Ce travail a bénéficié de deux collaborations et a donné lieu à deux publications: l'une dans la communauté de l'argumentation [1] et l'autre dans la communauté des treillis de concepts [2]. Plus récemment, j'ai étudié si les propriétés discursives des phrases sont intégrées dans leurs représentations distributionnelles, en construisant des tâches de classification spécialement conçues pour prédire les propriétés discursives de celles-ci. Ce travail préliminaire avec un stagiaire a donné lieu à une publication [3]. Nous y comparons les performances des approches contextuelles et non contextuelles des représentations de phrases pour détecter les propriétés discursives de celles-ci.

[1] Huber L, Toussaint Y, Roze C, Dargnat M, Braud C. Aligning Discourse and Argumentation Structures using Subtrees and Redescription Mining. ArgMining@ACL Proceedings of the 6th Workshop on Argument Mining, August 2019, Florence, Italy, 2019: 35-40

[2] Huber L, Reynaud J, Dargnat M, Toussaint Y. AOC-Poset on Discourse and Argumentation Subgraphs: What Can we Learn on Their Dependencies? CLA 2020: 107-118

[3] Huber L, Memmadi C, Dargnat M, Toussaint Y. Do sentence embeddings capture discourse properties of sentences from Scientific Abstracts ? CODI 2020 - EMNLP 1st Workshop on Computational Approaches to Discourse, Nov 2020, Punta Cana, Dominican Republic.

Thèse Détection généralisable des discours de haine en ligne avec adaptation du domaine et modélisation du sujet (2019-2022) - Tulika Bose

Sous la direction d'Irina Illina (LORIA - équipe Multispeech), Dominique Fohr (LORIA - équipe Multispeech) et Angeliki Monnier (CREM – équipe Pixel)

Nous étudions l'utilisation de l'identification automatique des sujets, au sens de thèmes, pour améliorer la détection des discours de haine. La recherche sur l'identification automatique des discours de haine dans les médias sociaux implique l'utilisation d'une variété de corpus dans la littérature, qui diffèrent en termes de sujets. Une analyse expérimentale détaillée a montré que la plus grande variété de sujets abordés dans les corpus de formation et le plus grand nombre de chevauchements thématiques entre les corpus de formation et de test permettent d'améliorer la généralisabilité des modèles de sujets («topic modeling») pour la détection des propos injurieux entre corpus. De plus, en raison d'un changement dans la distribution des différents corpus de discours haineux, nous étudions les mécanismes d'adaptation des domaines pour une meilleure généralisabilité des modèles formés.

WP1 - CONNAISSANCE ET INGÉNIERIE

Thèse Modélisation et inférence de la persistance de l'information sur les réseaux sociaux (2019-2022) - Nicolas Dante

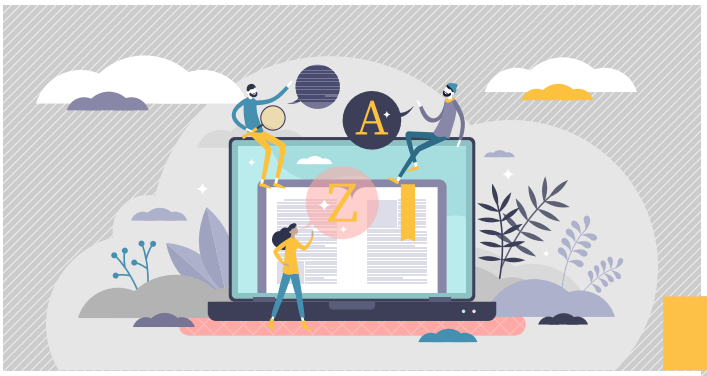
Sous la direction de Marianne Clausel (IECL– équipe Probstat) et Radu Stefan Stoica (IECL– équipe Probstat)

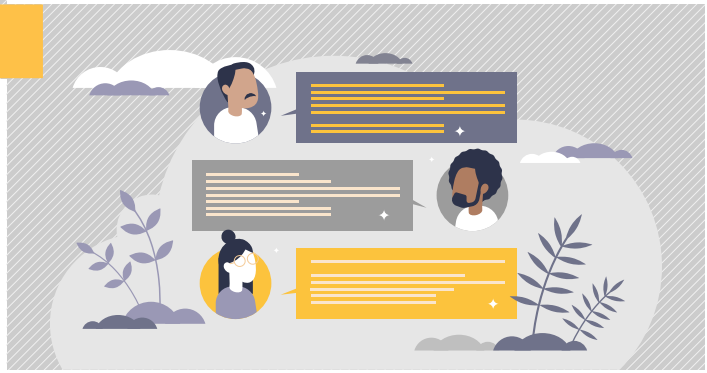
Afin de modéliser l'information contenue dans les réseaux sociaux, je me suis familiarisé avec la notion de «topic modeling». Il s'agit d'un ensemble de méthodes qui visent à représenter un corpus ou un texte à partir des thèmes sous-jacents. Je me suis particulièrement intéressé au modèle ATAM (Ailment Topic Aspect Model) conçu pour l'étude des tweets liés à la santé. Ce modèle permet la construction d'une série chronologique multivariée représentant l'évolution de la proportion des sujets. Je me suis ensuite consacré aux séries temporelles et plus particulièrement à la notion de «Longue Mémoire». Plusieurs méthodes d'inférence, basées sur le contenu spectral de la série temporelle multivariée, sont décrites dans la littérature. Je souhaite adapter une approche bayésienne d'estimation univariée du paramètre «Longue Mémoire» dans le cas multivarié.

Post-doc Données multi-sources et polymorphes : faire collaborer fouille de motifs et Formal Concept Analysis pour une meilleure extraction de connaissances

Jiajun Pan en collaboration avec Armelle Brun (LORIA- équipe KIWI) et Yannick Toussaint (LORIA- équipe Orpailleur)

Il est désormais classique de disposer de multiples sources de données portant sur un même phénomène ou des mêmes éléments. Un phénomène peut être un thème tel que les « fake news », la population animalière d'un pays, etc…ou de façon plus générale un domaine tel que la santé, le e-commerce, l'éducation, etc... Les données peuvent être des données textuelles, des données de description, des données de réalisation, etc…Les sources de données peuvent avoir des structures différentes et offrir des points de vue différents. Lorsque les sources ont des points de vue différents, chacune d'elles fournit a fortiori une connaissance différente et peut être complémentaire sur le phénomène. L'objectif du post-doc est de réaliser une fouille conjointe de données multi-sources afin d'en extraire une information plus riche et d'unifier la connaissance liée aux éléments étudiés. Nous nous intéressons ici à l'e-éducation avec des données de description d'activités pédagogiques d'apprenants, curriculum scolaire, etc…Le projet s'intéresse à la fois à la description d'éléments (au travers d'attributs) et à l'utilisation de ces éléments en contexte (en particulier dans un cadre séquentiel). La première approche envisagée est la fouille de redescriptions, qu'il faut adapter pour permettre de gérer les structures multiples des sources. Des approches telles que la fouille multi-vue ou la fouille de données relationnelles seront également explorées.





LA RECHERCHE

WP2 - LANGAGE

Mots-clés : ressources linguistiques, traitement automatique des langues, apprentissage des langues

La thématique principale du WP2 est le traitement automatique des langues. Il rassemble principalement des chercheurs issus des sciences du langage et de l'informatique. L'un des principaux objectifs est de consolider les collaborations entre ces deux disciplines.

Un premier aspect du WP2 concerne la constitution de ressources dans des domaines ciblés sous la forme de lexiques ou de corpus: ex. au niveau des lexiques, la constitution d'un dictionnaire des mots utilisés par les élèves de maternelle. Les corpus ont une place singulière pour cette problématique, ainsi leur constitution revêt un enjeu majeur. On trouve dans les travaux du work package: la construction d'un corpus d'exemples issus des dictionnaires, ou encore d'entretiens avec Emile Zola ; la constitution d'un corpus sur la chimie verte et son expression dans les réseaux sociaux; ou la constitution d'un corpus de transcription de parties du jeu de plateau les colons de Catane.

Un second aspect relève plus spécifiquement du traitement des données langagières comme par exemple la production d'études de modèles de la langue, en particulier pour la sémantique et son calcul (modèles sémantiques computationnels). Il s'agit d'étudier les plongements lexicaux et leurs propriétés afin d'extraire leurs caractéristiques par apprentissage automatique. Une autre question quant à la sémantique a été d'identifier les particularismes dans des dialogues, notamment entre des cliniciens et des patients souffrant de troubles mentaux. Cette focalisation sur des enjeux allant au-delà de la syntaxe montre que la communauté se projette sur des questions ambitieuses et difficiles.

Enfin, l'apprentissage humain de la langue est aussi une thématique forte, d'une part dans l'étude analytique du comportement des apprenants, en particulier en considérant les multi-modalités de l'environnement d'apprentissage, ainsi que la mise au point d'une plateforme d'aide à l'apprentissage en générant automatiquement des exercices adaptés au niveau de l'apprenant. Dans le WP2, les chercheurs s'expriment dans la diversité des thématiques et des problématiques aux frontières de l'informatique et de la linguistique.

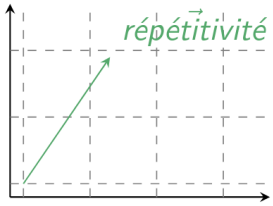
Thèse Génération automatique de définitions et de propriétés sémantiques de mots (2018-2021) - Timothee Mickus

Sous la direction de Mathieu Constant (ATILF – équipe Ressources) et Denis Paperno (Utrecht University, LORIA-équipe SyNaLP)

La génération automatique de définitions est une tâche du traitement automatique du langage (TAL) mettant en équivalence les représentations sémantiques distributionnelles (ou plongements lexicaux, « embeddings ») et les gloses de dictionnaires pour les entrées correspondantes. Le but de la thèse est de développer un cadre méthodologique solide afin de déterminer si ces deux types de représentations du sens sont équivalentes ou non. Parmi les aspects déjà abordés, nous pouvons citer :

1. une mise en rapport des architectures de génération de définition et des modèles plus communs de type « séquence-à-séquence » ;
2. une étude des différences d'architectures de plongements, et l'impact qui en découle sur les résultats obtenus sur la tâche de génération de définitions ;
3. une réflexion sur les limites des métriques de surface couramment utilisées en génération du langage (BLEU, perplexité) et en particulier sur leur inadaptation aux enjeux de la génération de définition, qui nécessite à présent la création d'un corpus d'annotation pour l'évaluation des lacunes des modèles distributionnels ;
4. une mesure de la similarité intrinsèque entre représentations vectorielles distributionnelles et représentations textuelles des gloses.

Les conclusions intermédiaires que nous tirons de l'état actuel de nos travaux suggèrent qu'une différence d'ancrage sémantique joue entre les représentations vectorielles distributionnelles (dont l'ancrage sémantique est très indirect) et les gloses de dictionnaires (qui sont écrites dans le but exprès de désigner le référent du mot à définir). Nous travaillons à présent à la consolidation de ces résultats préliminaires, notamment dans un contexte multilingue et diachronique.



Qualité, caractère de ce qui est répétitif (événements, gestes, etc.).

Thèse The Lexicon of the Environment and Green Chemistry in Ordinary Discourse. Using social networks as corpora – LEGCOD (2019-2022) - Tomara Gotkova

Sous la direction d'Alain Polguère (ATILF – équipe Lexique) et Francesca Ingrosso (LPCT)

La première année du projet LEGCOD a été consacrée à l'exploration du vocabulaire de l'environnement et de la chimie dans le discours de la langue générale trouvé sur Internet dans les réseaux sociaux ; plus précisément, dans le cadre de notre recherche : Twitter et Reddit. Nous avons commencé par construire un corpus spécialisé composé de textes scientifiques de l'environnement et de la chimie en anglais. En collaboration avec nos collègues de l'Université Pompeu Fabra à Barcelone nous avons utilisé ce corpus pour l'extraction automatique des termes clés de l'environnement et de la chimie. Le résultat a été utilisé comme un filtre pour construire un corpus de référence de données issues de Twitter et Reddit. Après avoir développé une technique d'extraction pour chaque réseau social, nous avons collecté des données préliminaires pour un corpus d'essai. L'étape suivante, qui sera réalisée au cours de la deuxième année, consistera à nettoyer et tester notre corpus d'essai pour l'analyse linguistique plus approfondie et l'étude du lexique de l'environnement et de la chimie dans le discours de la langue générale.

Post-doc Amélioration, expérimentation et analyse du dispositif numérique d'apprentissage du français FLEURON - Biagio Ursi

En collaboration avec Virginie André (ATILF – équipe Didactique des langues et sociolinguistique) et Manuel Rebuschi (AHP-PReST)

La base de données FLEURON (Français Langue Etrangère Universitaire Ressources et Outils Numériques) [1] propose des ressources multimédias authentiques, classées par catégories, qui illustrent un ensemble de situations de communication de la vie des étudiants en France. Ces ressources permettent de découvrir des situations auxquelles les étudiants sont confrontés dès leur arrivée dans une université française, elles permettent également d'observer différentes actions et interactions dans des situations variées de la vie universitaire et de tous les jours.

Le post-doc a contribué à l'alimentation de la base de données en prenant part à la collecte de nouveaux enregistrements audiovisuels ; il a participé à la conception d'expérimentations portant sur l'exploitation du concordancier de la plateforme, qu'il a menées avec des apprenants allophones inscrits aux cours de Français Langue Etrangère du Centre de Langue Yves Châlon de l'Université de Lorraine, selon différentes configurations : seuls et en binôme, avec l'accompagnement d'un enseignant ou en autonomie. Ces expérimentations ont été vidéo-enregistrées grâce à la mise en place d'un dispositif de captation adapté ; elles ont fait l'objet d'études qualitatives et d'évaluations épistémologiques qui ont donné lieu à plusieurs contributions dans des colloques internationaux [2] et articles scientifiques en cours de publication [3] ou de rédaction. Suite à ce post-doc, la collaboration se poursuit en vue de l'implémentation d'annotations de phénomènes conversationnels pour l'apprentissage du français parlé en interaction, à partir des ressources FLEURON.

[1] https://fleuron.atilf.fr

[2] Biagio Ursi. Corpus-based resources in conversation: Learning with the multimodal concordancer of the FLEURON database. Teaching and Language Corpora Conference - TaLC2020, Henry Tyne (chair), Jul 2020, Perpignan, France
Biagio Ursi, Virginie André. Corpus : exploration, médiation et autonomisation. L'utilisation du concordancier de la plateforme FLEURON en classe de FLE. La linguistique appliquée à l'ère digitale - Colloque VALS/ASLA 2020, Alain Kamber; Simona Pekarek Doehler; Maud Dubois (chairs), Feb 2020, Neuchâtel, Suisse

[3] Biagio Ursi, Virginie André. Corpus : exploration, médiation et autonomisation. L'utilisation du concordancier de la plateforme FLEURON en cours de FLE. Bulletin suisse de Linguistique appliquée, A paraître

LA RECHERCHE

WP3 - DÉFIS SOCIETAUX

Mots-clés : ressources linguistiques, épistémologie, éthique, humanités numériques, citoyens

Le WP3 concerne les défis sociétaux liés au traitement des données langagières et au recours à l'IA. Les objectifs sont d'une part de développer des corpus en sciences humaines et sociales permettant de procéder à des analyses de discours et de controverses, d'autre part d'approfondir les questions épistémologiques et éthiques liées aux usages des intelligences artificielles, notamment dans le champ scientifique.

Un premier ensemble de corpus a été construit à partir du réseau socionumérique Twitter. Ce sont les discours de haine contre les migrants qui en ont été la première application ; ces discours peuvent manifester la haine de leurs auteurs mais aussi inciter à des actes criminels envers les groupes de population visés. Un second ensemble de corpus concerne des discours relatifs à la transparence en matière d'environnement, de santé et de sécurité alimentaire circulant en France et en Europe. Il s'agit de corpus hétérogènes, constitués à la fois par exemple de textes institutionnels et d'articles de presse qui permettront l'analyse des controverses sur le sujet.

Les questions d'éthique liées au développement des outils numériques et de l'IA ont été abordées par le prisme du droit à l'explication, un sujet central en éthique de l'IA et par l'étude de données numériques de santé. Par ailleurs, une charte pour le respect des données privées (tant d'un point de vue technique que juridique) pour le développement des modèles de Machine Learning a été élaborée. Plus généralement, les recherches sur ces questions se sont inscrites dans une perspective interdisciplinaire entre politique, éthique et droit dans le numérique.

Les questions d'épistémologie liées à l'utilisation d'IA et de logiciels ont été approchées en se focalisant d'une part sur une étude de cas, à savoir l'histoire récente de la chimie computationnelle. Une analyse du corpus constitué par les échanges pendant une vingtaine d'années sur une liste de discussion de la communauté des chimistes computationnels montre les interactions fortes entre l'élaboration des modèles et des logiciels et fait apparaitre très tôt des débats méthodologiques, souvent tendus, autour des questions de transparence, de validité et de reproductibilité des méthodes de calcul. D'autre part, une autre direction est motivée par les pratiques innovantes en humanités numériques, à savoir l'application et le développement d'outils du Web sémantique pour des corpus de correspondance en histoire et philosophie des sciences.

Thèse Indexer et explorer un corpus d'humanités numériques par des représentations élastiques - Nicolas Lasolle

Sous la direction d'Olivier Bruneau (AHP-PReST) et Jean Lieber (LORIA, équipe K)

Ce projet s'intéresse à l'application et le développement d'outils du Web sémantique pour le corpus de la correspondance d'Henri Poincaré. Les travaux récents se sont concentrés sur la proposition de méthodes pour assister l'édition de données RDF. Un mécanisme utilisant le raisonnement à partir de cas et l'exploitation des connaissances de l'ontologie a été formalisé afin de fournir une liste de suggestions lors de l'édition d'un fait (triplet RDF). Un outil utilisant ces méthodes a été développé et testé avec le corpus de la correspondance d'Henri Poincaré. Ce travail, qui peut être réutilisé dans d'autres contextes, a été présenté lors de la conférence internationale du Web sémantique (ISWC 2020). Un autre travail en cours concerne un problème de représentation de connaissances lié à l'intégration de données temporelles lors de l'indexation de corpus historiques par des technologies du Web sémantique.

Post-doc Transparency discourses: from institutions to citizens (DISTIC) - Jana Vargovčíková

En collaboration avec François Allard-Huver, Anne Piponnier, Emmanuelle Simon (CREM, équipe PRAXIS) et Marianne Clausel (IECL)

Le sujet porte sur les discours relatifs à la transparence en matière d'environnement, de santé et de sécurité alimentaire circulant en France et en Europe.

Le projet DISTIC (Transparency Discourses: From Institutions to Citizens) s'intéresse aux significations multiples et parfois contradictoires du terme transparence dans les controverses liées à l'environnement, à la santé et à la sécurité alimentaire. En effet, la question de la disponibilité publique d'informations et d'expertises fiables se retrouve souvent au centre des débats sur les risques environnementaux ou sanitaires, comme le démontre d'ailleurs la crise actuelle due à l'épidémie de Covid-19.

Dans ce contexte, les discours sur la transparence produits par les autorités publiques, les acteurs industriels, les organisations non gouvernementales et les médias cristallisent les tensions entre, d'une part, l'accès public à de plus en plus d'informations et de données et, d'autre part, la défiance citoyenne croissante envers ces données et les savoirs dits experts qui les façonnent. La tâche est donc d'abord de construire et analyser des corpus de textes, pour mettre en lumière la circulation des discours sur la transparence entre différents espaces et types d'acteurs. Ensuite, nous pointerons les divergences dans les acceptions du terme transparence et des objectifs et limites inhérents chez ces différents acteurs. Pour cela, nous nous appuierons notamment sur des méthodes lexicométriques et sur des méthodes qualitatives d'analyse des discours et nous nous concentrerons sur les terrains français et européen. En plus de contribuer à l'accroissement des connaissances scientifiques, cette recherche pourra aider les acteurs de la société civile et les citoyens pour se repérer dans les injonctions parfois contradictoires à la transparence, ainsi que pour réfléchir aux présupposés de leurs propres appels à plus de visibilité sur les politiques publiques.

Post-doc Ethics of AI and IT - Maël Pégny

En collaboration avec Anna Zielinska (AHP-PReST), Cyrille Imbert (AHP-PReST) et Christophe Cerisara (LORIA – équipe SyNaLP)

Le début du post-doc a été consacré à l'achèvement d'un article sur le droit à l'explication, un sujet classique de la littérature en éthique de l'IA [1].
La principale tâche du post-doc consistait en la rédaction d'une charte sur l'éthique de l'IA, fondée sur une collaboration entre les Archives Henri Poincaré et le LORIA. Nous avons centré notre travail sur le développement des modèles de *machine learning* respectueux de la vie privée dès la conception. Le sujet présentait deux avantages tactiques importants. Le premier était sa pertinence pour l'équipe de Traitement Automatique de la Langue très impliquée dans le projet, car l'apprentissage de modèles sur de vastes corpus de textes écrits pose des problèmes évidents de respect de la vie privée. Le second était le faible développement de la littérature sur les attaques par inversion de modèles. Ces dernières consistent à récupérer les données encodées dans les modèles, et peuvent être opérées même lorsque les données d'apprentissage ont été détruites : elles constituent donc un enjeu de respect de la vie privée nouveau et spécifique à l'IA. Le document achevé comprend dix recommandations concrètes aux développeurs, ainsi qu'une discussion des frontières de l'état de l'art sur ces questions de respect de la vie privée tant du point de vue technique, comme les problèmes de reconnaissance d'informations privées dans les corpus, et juridiques, comme les questions de portée de la définition des données personnelles. Le document invite un retour d'expérience de la part des développeurs qui pourra servir de base à des travaux futurs.

Par ailleurs, nous avons publié un bref article sur les données médicales, et leur place dans les mouvements sociaux dans l'hôpital public. Cet article se plaçait sur le double front de la recherche académique et de sa diffusion auprès du public, afin de pouvoir contribuer aux débats en cours sur l'avenir du système de santé. L'article a été fondé à la fois sur la littérature académique et journalistique et sur des interviews de soignants impliqués dans ces mouvements sociaux [2].

[1] Pégny M, Thelisson E, Ibnouhsein I. The Right to an Explanation Delphi - Interdisciplinary Review of Emerging Technologies. (2020) Volume 2, Issue 4 pp. 161 - 166. https://doi.org/10.21552/delphi/2019/4/5

[2] Pégny M, Zielinska A. L'épineuse question des données numériques de santé. (2020) https://theconversation.com/lepineuse-question-des-donnees-numeriques-de-sante-131586

Post-doc Discours haineux en ligne contre les migrants - Axel Boursier et Nadia Makouar

En collaboration avec Angeliki Monnier (CREM – équipe Pixel), Irina Illina (LORIA – équipe Multispeech) et Dominique Fohr (LORIA– équipe Multispeech)

La mise en œuvre de processus automatisés de détection des discours de haine sur Internet (médias socio-numériques, sections de commentaires dans les journaux, etc.), nécessite une meilleure compréhension du discours de haine en tant que phénomène social. Pour cette raison, les deux chercheurs postdoctoraux en sciences humaines et sociales se sont penchés sur les enjeux de l'émergence et de la circulation des discours de haine en ligne, ainsi que sur la structure discursive de ceux-ci. Les travaux ont porté sur le cas précis du discours de haine en ligne contre les migrants. Plusieurs publications et communications ont permis la publicisation de leurs travaux [1-3], dont une sélection figure ci-dessous :

[1] A. Boursier, « Circulation des discours de haine dans la sphère publique numérique », 3rd DiscourseNet Congress « Language and power in a polycentric world », Université de Cergy-Pontoise, 12 septembre 2019 (publication en cours dans l'Harmattan, collection « Cahiers de la nouvelle Europe »).

[2] A. Boursier "Media truth is not mine", séminaire University of Warwick, 13 novembre 2019.

[3] N. Makouar, "Anti-intellectualism as a strategy of anti-immigrant propaganda: semantic analysis of French media online comments regarding immigration statistics", 24th DiscourseNet Conference "Discourse and Communication as propaganda: digital and multimodal forms of activism, persuasion and disinformation across ideologies", Bruxelles, Septembre 2020.

ENVIRONNEMENT DU PROJET

Le projet OLKi renforce une dynamique locale autour de questions liées à l'éthique et à l'apprentissage automatique

Projet H2020 AI-PROFICIENT
(Artificial Intelligence for Improved PROduction efFICIEncy, quality and maiNTenance)
CRAN & LORIA

Projet H2020 COMPRISE
(Cost-Effective, Multilingual, Privacy-Driven Voice-Enabled Services)
LORIA

Réseau de formation H2020 NL4XAI
(Natural Language for Explainable AI)
LORIA

Projet ANR DFG M-PHASIS
(Migration and Patterns of Hate Speech in Social Media - A Cross-cultural Perspective)
CREM & LORIA

Projet IA ANR-DFG-JST EDDA
(EnhanceD Data stream Analysis with the Signature Method)
IECL & autres partenaires

Projet ANR QUANTUM
(Génération de questions pour la compréhension textuelle par lecture automatique)
IRIT & Loria & Synapse

Chaire IA XNLG
(Generating Text in Multiple Languages from Multiple Sources)
LORIA

Séminaire philosophie - informatique
AHP & LORIA

Projet E-FRAN METAL
(Modèles Et Traces au service de l'Apprentissage des Langues)

Atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)
LORIA

Projet MSH LogiScienS
(Science ouverte et logiciels : une histoire de la chimie computationnelle au prisme d'une biographie de logiciels scientifiques)
AHP, ATILF, LORIA & autres partenaires

Équipement d'excellence ORTOLANG
(Outils et Ressources pour un Traitement Optimisé de la LANGue)
ATILF

Base de données FLEURON
(Français Langue Étrangère Universitaire Ressources et Outils Numériques)
ATILF

Collaboration sur la fouille de publications scientifiques
IECL, ATILF & Cancéropôle Est

LA FORMATION

Écoles thématiques

École d'été Python for Natural Language Processing (Python4NLP) - 26 - 30 août 2019

Pour la première édition de l'école, 46 personnes, de 9 pays et de 25 institutions différentes, sont venues du 26 au 30 août 2019 au laboratoire LORIA.
Cette école s'adressait aux étudiants, chercheurs et autres personnes intéressées à la fois par l'informatique et les humanités (en particulier la linguistique). L'école visait à développer des compétences en collecte, traitement et analyse de données textuelles en s'appuyant sur le langage de programmation Python enrichi des librairies dédiées au traitement textuel. En plus des cours théoriques, le programme proposait des conférences invitées, des travaux pratiques et des événements sociaux.
Les conférenciers invités étaient Albert Gatt (Université de Malte) et Malvina Nissim (Université de Groningen, Pays-Bas). Les cours ont été dispensés par Chloé Braud, Christophe Cerisara, Claire Gardent et Yannick Parmentier de l'équipe Synalp (LORIA) et les travaux pratiques ont été encadrés par Thien Hoa Le, Esteban Marquer, Timothée Mickus, Siyana Pavlova, et Anastasia Shimorina (étudiants du Master TAL ou doctorants au LORIA).
Le retour des apprenants est extrêmement positif ; 91% d'entre eux se déclarant satisfaits ou très satisfaits de l'école.

L'école a été organisée et financée par l'initiative Lorraine Université d'Excellence (programme RECOLTE), le projet IMPACT LUE Open Language and Knowledge for Citizens (OLKi) et le Groupe d'Intérêt de Recherche Linguistique Informatique Formelle et de Terrain (GDR CNRS LIFT).

École d'automne AI & education - 17- 25 octobre 2019

Cette école a accueilli 19 participants, à la fois étudiants de Master, doctorants et chercheurs.
La thématique principale était l'Intelligence Artificielle et l'éducation, et visait à comprendre les défis liés à l'intégration de l'Intelligence Artificielle dans le monde de l'éducation qui subit actuellement des changements profonds.
Cette école était voulue très pluridisciplinaire.
Elle s'est déroulée en 3 temps :
- du 17 au 21 octobre, des sessions de cours ont eu lieu, avec des participants traitant de la fouille de données, de la constitution automatique de tests, la gestion des connaissances, la science des données centrée-humain, etc.
- du 22 au 23 octobre, les participants ont pu assister à la conférence LSAC (Learning and Students Analytics Conférence) en assistant à des présentations de chercheurs, entreprises et acteurs de l'éducation.

- du 24 au 25 octobre, les participants ont participé à un datathon de deux jours ayant pour but d'appliquer les éléments vus les jours précédents sur un jeu de données concret.

L'école a été organisée et financée par l'initiative Lorraine Université d'Excellence (programme RECOLTE), le projet IMPACT LUE Open Language and Knowledge for Citizens (OLKi) et l'Institut des Sciences du Digital, Management et Cognition (IDMC).

TEMPS FORTS

2018

OCTOBRE

Démarrage du projet OLKi

OCTOBRE

Démarrage de 2 thèses

26-30 AOÛT

École d'été en traitement automatique des langues Python4NLP

23 AVRIL

Séminaire technique Intelligence Artificielle

19 SEPTEMBRE

1ère session du Séminaire philosophie - informatique L'ontologie appliquée au service du partage des données par Adrien Barton

OCTOBRE

Démarrage de 4 thèses

2019

6 FÉVRIER

Meetup IA: l'intelligence artificielle appliquée au traitement du langage naturel

28 MARS

Présentation du projet OLKi aux "Journées AMI Grand Débat"- CNRS

7 MARS

« Le projet OLKi : vers une IA citoyenne et ouverte », sensibilisation d'enseignants du secondaire lors de la Journée Informatique et Science du Numérique

15 MARS

Journée IA, Langage et Citoyens

16-17 OCTOBRE

International Conference on Arabic Language Processing (ICALP 2019)

TEMPS FORTS

17-25 OCTOBRE

École d'automne Artificial Intelligence and Education

19 NOVEMBRE

Journée d'étude «Les pratiques journalistiques face aux algorithmes et l'automation»

28 NOVEMBRE

Petit-déjeuner IA : TAL & Big Data rencontre entre scientifiques et acteurs économiques

2020

8 FÉVRIER

Présentation du projet OLKi à la session grand public de l'événement A.I_Now

6 MARS

Intervention à la journée spéciale «La place des femmes dans le numérique» Session - Pourquoi les femmes doivent s'intéresser à l'IA?

8-12 JUIN

Conférence JEP-TALN-RECITAL 2020

13-14 OCTOBRE

Journées Omeka Colloque Sciences Ouvertes Prendre soin de ses données et les valoriser

28-29 SEPTEMBRE

Conférence Document numérique et société - Humains et données : création, médiation, décision, narration

8 SEPTEMBRE

Matinée d'étude sur les données issues des réseaux sociaux

À venir...

Février 2021 : point d'étape du projet

Workshop «Éthique, droit et IA»

Été 2021 : École thématique «Corpus et didactique»

Été 2021 : École Python4NLP - 2ème édition

CONTACTS

Comité opérationnel

François ALLARD-HUVER (CREM), MCF UL
Maxime AMBLARD (LORIA), MCF UL, co-leader du WP 2 Langage
Christophe CERISARA (LORIA), CR CNRS, porteur scientifique du projet OLKi
Maud CIEKANSKI (ATILF), MCF UL
Marianne CLAUSEL (IECL), Professeure UL, co-leader du WP 1 Connaissance et ingénierie
Aurore COINCE, IR UL, chargée de projet
Dario COMPAGNO (CREM), MCF UL
Mathieu CONSTANT (ATILF), Professeur UL, co-leader du WP 2 Langage
Philippe NABONNAND (AHP-PReST), Dir AHP-PReST, co-leader du WP 3 Défis sociétaux
Brigitte SIMONNOT (CREM), Professeure UL, co-leader du WP 3 Défis sociétaux
Radu-Stefan STOICA (IECL), Professeur UL
Yannick TOUSSAINT (LORIA), Professeur UL, co-leader du WP 1 Connaissance et ingénierie
Pierre WILLAIME (AHP-PReST), IE CNRS

Membres invités

Xavier ANTOINE, Directeur de l'IECL (jusqu'en août 2020)
Alex BOULTON, Directeur de l'ATILF
Thierry DAUNOIS, représentant de la DEPAS (à partir de 2020)
Kevin DEGIORGIO, représentant de la DRV
Anne GEGOUT-PETIT, Directrice de l'IECL (à partir de septembre 2020)
Inga GIRARDIN, représentante de la Satt Sayens (à partir de 2020)
Annette JARLEGAN, Directrice du Pôle CLCS
Yves LAPRIE, Directeur du Pôle AM2I
Jean-Yves MARION, Directeur du LORIA
Jacques WALTER, Directeur du CREM
Nathalie WITTMANN, représentante de la DEPAS (jusque fin 2019)

Projet IMPACT OLKi

Laboratoire lorrain de recherche en informatique et ses applications
Campus scientifique
BP 239
54506 Vandoeuvre-lès-Nancy Cedex
Tél : +33 3 83 59 20 00

christophe.cerisara@loria.fr
aurore.coince@univ-lorraine.fr

lue.univ-lorraine.fr/open-language-and-knowledge-citizens-olki

PARTENAIRES

5 laboratoires

Archives Henri-Poincaré
Philosophie et Recherches sur les Sciences et les Technologies

atilf

crem centre de recherche
EA 3476 sur les médiations
communication, langue, art, culture

Institut ÉLIE CARTAN

Loria

Un projet Lorraine Université d'Excellence

INVESTIR L'AVENIR

UNIVERSITÉ DE LORRAINE

LORRAINE Université d'Excellence

IMPACT OLKi

Partenaires LUE

UNIVERSITÉ DE LORRAINE

cnrs

CHRU NANCY

INRAE

Inserm

Inria

Georgia Tech Lorraine

AgroParisTech

Publication : janvier 2021
Conception - réalisation : Loria

Crédit photos : OLKi, membres et partenaires du projet OLKi, Freepik, Undraw, Flaticon