3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic: Decentralized coordination of intelligent processes to improve the quality of a collection of knowledge graphs
- Research axis of the 3IA: Axis 1 - Core Elements of AI
- **Supervisor (name, affiliation, email): Fabien Gandon, Centre Inria d'Université Côte d'Azur, fabien.gandon@inria.fr**
- Potential co-supervisor (name, affiliation): Pierre Monnin, Université Côte d'Azur
- The laboratory and/or research group: WIMMICS team, Centre Inria d'Université Côte d'Azur

---

Apply by sending an email directly to the supervisor.
The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master's degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

- Description of the topic:

This thesis aims to study how different agents can manipulate and refine knowledge graphs by leveraging different forms of intelligent processes (e.g., deductive, analogical, inductive reasoning) while collaborating before, during and after the execution of their respective processes. The challenge of such an approach resides in being able to mobilize and combine, with low software dependency, as many artificial intelligence methods as necessary to improve the quality and quantity of knowledge available in a collection of sources.

(detailed subject in subsequent pages)

(FR) Coordination décentralisée de traitements intelligents pour améliorer la qualité d'une collection de graphes de connaissances
(EN) Decentralized coordination of intelligent processes to improve the quality of a collection of knowledge graphs

Fabien Gandon, Pierre Monnin
Equipe Wimmics
(Université Côte d'Azur, Inria, CNRS, I3S)

Mars 2024

Résumé

(FR) Cette thèse vise à étudier comment différents agents peuvent manipuler et améliorer des graphes de connaissances en employant différentes formes de traitements intelligents (par exemple, raisonnement déductif, analogique, inductif) tout en collaborant avant, au cours, et après la mise en oeuvre de leurs méthodes respectives. L'enjeu d'une telle approche est de pouvoir mobiliser et combiner, avec un faible couplage logiciel, autant de méthodes d'intelligence artificielle que nécessaire pour améliorer la qualité et la quantité de connaissances disponibles dans une collection de sources.

(EN) This thesis aims to study how different agents can manipulate and refine knowledge graphs by leveraging different forms of intelligent processes (*e.g.*, deductive, analogical, inductive reasoning) while collaborating before, during and after the execution of their respective processes. The challenge of such an approach resides in being able to mobilize and combine, with low software dependency, as many artificial intelligence methods as necessary to improve the quality and quantity of knowledge available in a collection of sources.

1 Contexte de la thèse

Cette thèse s'intéresse aux graphes de connaissances (GCs), artefacts ayant pour objectif de capturer formellement les connaissances de mondes modélisés,

leurs nœuds représentant des entités d'intérêt et leurs arêtes des relations entre ces entités [9]. Ils sont par exemple utilisés pour l'intégration de données de sources multiples, la gestion de données distribuées, ou pour représenter des réseaux complexes dans de multiples domaines comme la biologie, la sociologie, ou les processus industriels. En particulier, nous considérons les GCs représentés avec les formalismes du Web Sémantique et les principes des Données Liées.

Au sein du "Web des données" ou "Web Sémantique" [2], ces GCs sont représentés avec les langages RDF, SHACL, RDFS, et OWL standardisés par le World Wide Web Consortium (W3C), et sont interprétables à la fois par des humains et des agents logiciels, notamment des systèmes d'IA. En effet, les différents profils de OWL [8] sont basés sur des logiques de description [3], permettant notamment à des agents IA l'utilisation de mécanismes de raisonnement déductif dans l'exploitation des GCs. RDFS supporte des raisonnements plus légers alors que SHACL est un langage de validation de contraintes structurelles. Et d'autres langages existent pour représenter des règles d'inférence au-dessus de RDF par exemple en étendant les langages SPARQL ou SHACL. Ces dernières années, de nombreux modèles d'IA non-symbolique ont également été proposés, par exemple les modèles de plongement de graphes [1] et le *graph machine learning* [15]. Ceux-ci manipulent avec performance les GCs et ont une grande capacité à faire face à l'hétérogéneité et au bruit inhérents à leurs processus de construction (semi-)automatiques ou en *crowdsourcing*. Dans une perspective d'IA neuro-symbolique, plusieurs auteurs ont étudié l'injection des connaissances symboliques des GCs dans de tels modèles afin d'améliorer leur performance [6, 10, 14]. Plus récemment, des travaux encore peu nombreux se sont intéressés à l'utilisation du raisonnement par analogie pour les tâches du cycle de vie des GCs [11, 18] et ses interactions avec les modèles de plongement existants [13].

2 Combiner différents traitements intelligents

Les GCs peuvent donc d'ores et déjà être manipulés par plusieurs types de traitements intelligents, ici l'apprentissage, le raisonnement déductif, la validation de contraintes et le raisonnement analogique. Néanmoins, actuellement, rien ne permet de facilement combiner différentes méthodes d'IA de façon synergique pour une tâche donnée. Pourtant nous, humains, combinons tous les jours ces facultés dans des tâches complexes. De plus, les entrées-sorties et les objectifs des processus intelligents peuvent être naturellement composables. Ainsi un processus d'induction pourra produire des contraintes d'intégrité (ex. en SHACL) qui seront en suite utilisées par un processus de validation et mais aussi par un autre processus de complétion de graphes. C'est pourquoi, nous proposons dans le cadre de cette thèse d'étudier comment manipuler des GCs avec (1) différentes formes de traitements intelligents tout en (2) favorisant leurs interactions en les implémentant sous forme d'agents autonomes.

(1) Différentes formes de traitements intelligents. Afin de progresser vers une synergie entre types de processus intelligents dans une optique d'IA neuro-symbolique, nous étudierons comment manipuler et améliorer des GCs avec différentes formes de traitements intelligents (par exemple, raisonnement déductif, analogique, inductif). Nous considérerons les tâches d'amélioration de GCs [16] : prédiction de liens, classification de liens, alignement d'entités. Il sera nécessaire d'étudier les différents traitements intelligents mis en œuvre dans la littérature pour leur réalisation, leurs avantages, et leurs inconvénients. Cette étude permettra de pouvoir proposer de nouvelles approches pour la réalisation d'une tâche en utilisant un type de traitement intelligent actuellement non-considéré. Il sera également possible d'étudier d'autres tâches d'amélioration des GCs (par exemple, l'élagage [11]) ou d'autres processus intelligent (par exemple, raisonnement plausible), en fonction des appétences du doctorant.

(2) Collaboration d'agents intelligents. Dans l'optique d'un faible couplage logiciel, chaque processus intelligent sera considéré comme un "module" indépendant, capable d'interagir avec des GCs pour mener à bien des tâches. Une telle manipulation sera facilitée par la flexibilité de la pile des standards des GCs qui permet de créer des piles alternatives pour travailler avec différentes vues (par exemple, vue graphe, vue logique de descriptions, vue règles d'inférences, vue contraintes d'intégrité, etc.) ou avec différents processus intelligents. En particulier, nous proposons d'étudier cette organisation sous la forme d'un système multi-agents (SMA), une architecture d'IA distribuée explicitement conçue pour organiser et faire collaborer des composants logiciels d'IA faiblement couplés et pouvant être rendue compatible avec l'architecture décentralisée du Web des Données [5]. Ainsi une possibilité sera d'aller vers des architectures de type hMAS (Hypermedia Multi-Agent System) [4], une architecture de systèmes multi-agents hautement compatible avec l'architecture Web. En bénéficiant de cette architecture multi-agents, nous proposerons un cadre pour une interaction entre modules / processus intelligents avant, pendant, et après leurs processus respectifs, afin de bénéficier des apports de chaque type pour améliorer globalement la qualité d'une collection de graphes de connaissances.

Il est à noter que ce travail s'insèrera aussi dans la tendance actuelle à représenter et à négocier dans des GCs, sous forme de méta-données, les profils de raisonnements, et plus généralement les traitements, appliqués aux GCs publiés sur le Web. Dans cette vision, les GCs constitueront *(i)* le matériel en entrée des agents intelligents, *(ii)* la structure d'échange de leurs inférences (préliminaires, partielles, ou finales) leur permettant d'échanger et de consolider leurs résultats respectifs si nécessaire au cours de leur exécution, mais aussi *(iii)* le rôle de métadonnées pour décrire l'environnement dans lequel les agents évoluent évoluent, ses ressources, ses organisations et les autres agents. Les langages de représentation des GCs pourront donc être étendus pour représenter et décrire les traitements intelligents effectués au sein même des graphes (sur le modèle du langage OWL et du raisonnement déductif), incluant par exemple

leurs paramètres, configurations, traitements, workflows, et protocoles d'interaction. Les GCs constitueront ainsi une structure unifiée et pivot pour représenter connaissances et agents intelligents, dans la continuité des travaux existants pour représenter la provenance et tracer les traitements sur le Web des Données [12].

3 Étapes prévues dans le plan de travail

Cette thèse commencera par un état de l'art des différents domaines concernés notamment : la représentation des connaissances à base de graphes, les différentes familles de méthodes d'intelligence artificielle pour le traitement de GCs et les différentes approches de coordination dans les systèmes multi-agents.

Un premier phasage du travail de thèse pourra consister à étudier un nombre restreint et choisi de traitements intelligents combinables pour un sous ensemble de combinaisons intéressantes (ex. induction et dérivation logique).

Un deuxième phasage possible de la thèse sera d'étudier dans un premier temps la collaboration de plusieurs agents intelligents au-dessus d'un seul et même GC puis, dans un deuxième temps de considérer une collaboration impliquant plusieurs GCs par exemple répertoriés dans un catalogue de données.

Enfin, le plan de travail de cette thèse pourra aussi être structuré par l'exploration des différents mécanismes de mise en place de la collaboration des agents (différents protocoles comme dans FIPA [17], différentes organisations [19], orchestration vs chorégraphie vs stigmergie [7], etc). Une question de recherche pourra ainsi être la comparaison de ces différents mécanismes de collaboration ou, au contraire, leur combinaison.

4 Interactions prévues

Cette thèse bénéficiera d'interactions avec des projets et collaborations existantes de l'équipe Wimmics :

- Projet ANR AT2TA ("Analogies : from Theory to Tools and Applications")[1] : projet national français ayant pour objet d'étude le raisonnement par analogie et ses applications à différents types d'objets (textes, codes sources, dossiers patients, graphes de connaissances).
- Collaboration scientifique avec Claudia d'Amato[2] (Università degli Studi di Bari Aldo Moro – Italie) : collaboration scientifique autour de l'injection de connaissances symboliques dans les modèles de machine learning pour manipuler les GCs.
- Projet Franco-Suisse HyperAgents avec la collaboration de l'Université de St Gallen et l'Ecole des Mines de St Etienne, portant sur l'architecture hMAS pour la collaboration d'agents intelligents sur le Web.

1. https://at2ta.loria.fr/
2. http://www.di.uniba.it/~cdamato/

5 Contacts

Fabien Gandon ✉ fabien.gandon@inria.fr http://fabien.info
Pierre Monnin ✉ pierre.monnin@inria.fr https://pmonnin.github.io

Références

[1] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. Bringing light into the dark : A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12) :8825–8845, 2022.

[2] Dean Allemang, Jim Hendler, and Fabien Gandon. *Semantic web for the working ontologist : Effective modeling for linked data, RDFS, and OWL*. ACM, 2020.

[3] Franz Baader. *The description logic handbook : Theory, implementation and applications*. Cambridge university press, 2003.

[4] Andrei Ciortea, Olivier Boissier, and Alessandro Ricci. Engineering world-wide multi-agent systems with hypermedia. In *Engineering Multi-Agent Systems : 6th International Workshop, EMAS 2018, Stockholm, Sweden, July 14-15, 2018, Revised Selected Papers 6*, pages 285–301. Springer, 2019.

[5] Andrei Ciortea, Simon Mayer, Fabien Gandon, Olivier Boissier, Alessandro Ricci, and Antoine Zimmermann. A Decade in Hindsight : The Missing Bridge Between Multi-Agent Systems and the World Wide Web. In *AAMAS 2019 - 18th International Conference on Autonomous Agents and Multiagent Systems*, page 5, Montréal, Canada, May 2019.

[6] Claudia d'Amato, Nicola Flavio Quatraro, and Nicola Fanizzi. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, volume 12731 of *Lecture Notes in Computer Science*, pages 441–457. Springer, 2021.

[7] Marco Dorigo, Eric Bonabeau, and Guy Theraulaz. Ant algorithms and stigmergy. *Future generation computer systems*, 16(8) :851–871, 2000.

[8] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, Sebastian Rudolph, et al. Owl 2 web ontology language primer. *W3C recommendation*, 27(1) :123, 2009.

[9] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021.

[10] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. Treat different negatives differently : Enriching loss functions with domain and range constraints for link prediction. In *The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26 - 30, 2024, Proceedings*.

[11] Lucas Jarnac, Miguel Couceiro, and Pierre Monnin. Relevant entity selection : Knowledge graph bootstrapping via zero-shot analogical pruning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 934–944. ACM, 2023.

[12] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o : The prov ontology. *W3C recommendation*, 30, 2013.

[13] Hanxiao Liu, Yuexin Wu, and Yiming Yang. Analogical inference for multi-relational embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2168–2178. PMLR, 2017.

[14] Pierre Monnin, Chedy Raïssi, Amedeo Napoli, and Adrien Coulet. Discovering alignment relations with graph convolutional networks : A biomedical case study. *Semantic Web*, 13(3) :379–398, 2022.

[15] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proc. IEEE*, 104(1) :11–33, 2016.

[16] Heiko Paulheim. Knowledge graph refinement : A survey of approaches and evaluation methods. *Semantic Web*, 8(3) :489–508, 2017.

[17] Stefan Poslad. Specifying protocols for multi-agent systems interaction. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 2(4) :15–es, 2007.

[18] Zhen Yao, Wen Zhang, Mingyang Chen, Yufeng Huang, Yi Yang, and Huajun Chen. Analogical inference enhanced knowledge graph embedding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4801–4808. AAAI Press, 2023.

[19] Franco Zambonelli, Nicholas R Jennings, and Michael Wooldridge. Organisational abstractions for the analysis and design of multi-agent systems. In *Agent-Oriented Software Engineering : First International Workshop, AOSE 2000 Limerick, Ireland, June 10, 2000 Revised Papers 1*, pages 235–251. Springer, 2001.