DOING@MADICS

Bilan 2022 (version détaillée)

Mirian Halfeld Ferrari Alves (LIFO)
Anne-Lyse Minard Forst (LLL)
Genoveva Vargas-Solar (LIRIS)

1 Introduction

DOING aborde l'exploitation intelligente, efficace et sûre des documents par une recherche interdisciplinaire surpassant une simple mise à disposition de données. Deux enjeux de la thématique guident les réflexions proposées par notre action :

- La transformation des données en information.
- La transformation de l'information en connaissances.

La réflexion et le travail sur ces enjeux sont faits dans une perspective multi-disciplinaire avec notamment les domaines du *traitement automatique des langues, des bases de données et de l'intelligence artificielle (analyse de graphes)*. Notre réflexion est, jusqu'à maintenant, associée au domaine d'application de la santé.

Nous rappellons que DOING@MADICS est né d'une extension nationale du groupe de travail régional DOING@DIAMS qui motive des collaborations au sein de la Région Centre Val du Loire.

2 Activités réalisées

L'année 2022 a été en partie marquée par la crise sanitaire. Notre planning d'activités a ainsi été réalisé en mode hybride. Dans cette section, nous proposons un bilan organisé par type d'activité ; les activités sont décrites dans un ordre chronologique.

Nous avons adoptée une méthodologie de travail basée sur l'organisation de journées d'étude autour des thématiques précises à savoir la conception de bases de données à graphes à partir du contenu textuel, et l'utilisation de modèles d'intelligence artificielle pour analyser les graphes.

2.1 *Journées de travail hybrides*

Cette année nous avons organisé 3 journées de travail, dont une qui a eu lieu pendant le symposium du GDR MADICS à Lyon. Le programme des deux journées de travail incluait un keynote et une discussion sur le thème. L'objetif des temps d'échange était d'identifier des problématiques autour des bases de données graphes, leur conception, leur interrogation et leur analyse (i.e., data science queries). Les discussions dans les groupes de travail étaient riches et nous pensons reproduire ce format lors des prochaines rencontres DOING. Pendant le troisième webinar nous avons eu deux présentations invitées et un tutoriel.

2.2 Réunions

Nous avons organisé des réunions de travail et en prévoyons encore une en novembre 2022 (cf.infra).

- Séminaire de Nicolas Travers (ESILV - École d'Ingénieurs Généraliste Leonard da Vinci) au LIFO, suivi d'une discussion pour la mise en place de collaboration avec DOING.
- Journée de travail sur la conception de bases de données graphes pour représenter le contenu des cas cliniques avec Agata Savary (Unversité Paris Saclay - LISN). Nous avons échangé sur les problèmes associés à la conception des schémas de graphes de données adaptée à la représentation des textes, particulièrement, des cas cliniques.
- Réunion de travail (8 juin 2022) avec les responsables de l'atelier ROCED de MADICS pour préparer un panel conjoint organisé pour le symposium MADICS 2022, autour de la représentation des connaissances et le lien au texte.
- Visite de Martin Musicante de l'*Universidade Federal do Rio Grande do Norte*, Brésil, visite courte au LIRIS. Dans ce cadre, DOING organise un atelier de travail dans le but de définir des nouvelles étapes dans les recherches menées en collaboration avec Martin Musicante, qui est un collaborateur de longue date de Genoveva Vargas et Mirian Halfeld Ferrari.
- Semaine de travail DOING au LIFO (7 - 10 novembre 2022) : elle a porté sur des discussions et des sessions de travail concernant la modélisation de bases de données graphes à partir des exemples de cas cliniques et l'expression de requêtes data science sur les graphes. En particulier, la discussion s'est concentrée sur les langages déclaratifs, et les modèles IA pour l'analyse de graphes et leur modélisation sous forme d'opérateurs pour les intégrer à des requêtes Data Science. La semaine a inclus des sessions hybrides de présentations de travaux d'étudiant.e.s en thèse des différentes équipes.

2.2.1 Participation à la journée ARA

DOING a participé à la journée ARA organisée par MADICS en Mars 2022. Nous avons fait une présentation sur l'état d'avancement de nos refléxions par rapport aux verrous scientifiques que nous avons identifiés. Nous avons insisté sur la problématique de conception des bases de données graphes à partir du contenu textuel, et plus particulièrement des documents de spécialité comme les cas cliniques.

2.2.2 Fourth Symposium GDR CNRS MADICS : Journée de travail DOING

Dans le cadre du Symposium MADICS, DOING a organisé une demi-journée [1] avec un *Keynote* et un panel en collaboration avec l'atelier ROCED.

- 14:00 5 - Introduction et présentation de la demi-journée
- 14:10:45 – 14:55 - **Keynote : Aperçu général des langages de requêtes pour graphes à propriétés**, Victor Marsault, CNRS, Laboratoire d'Informatique Gaspard-Monge (LIGM, UMR 8049), Gustave Eiffel University & CNRS, Marne-la-Vallée, France
- 14:55 – 16:00 - **Panel : Les graphes dans la représentation, du contenu et de la connaissance dans les textes : les systèmes de gestion de graphes et les approches sémantiques**

  Panéliste :

  - Donatello Conte, Polytech Tours de l'Université de Tours, LIFAT
  - Nathalie Hernandez, Université de Toulouse - Jean Jaurès (ROCED)
  - Agata Savary, Université Paris-Saclay, LISN
  - Nicolas Travers, ESILV

  Moderateurs :

  - Mirian Halfeld Ferrari Alves, Université d'Orléans, LIFO
  - Catherine Roussey, INRAE (ROCED)
  - Genoveva Vargas-Solar, CNRS, LIRIS

  La discussion s'est organisée autour d'une question leitmotiv :

  Quelle est la place de la modélisation/interrogation du contenu des textes dans le traitement automatique des langues, le Web sémantique, la conception des BD pour les SGBD à Graphes, et l'apprentissage automatique/fouille de données ?

Un papier de positionnement est en cours de rédaction par les participant.e.s et les modératrices du panel.

[1] `https://www.univ-orleans.fr/lifo/evenements/doing/?page_id=678`

2.3 Les contacts et rayonement de l'action

Nous avons établie des contacts nationaux avec différentes personnes et groupes qui travaillent sur les graphes en bases de données, IA et du point de vue formel et qui s'intéressent au traitement du contenu textuel.

Dans un contexte international nous continuous à travailler avec l'Université Rio Grande do Norte et l'Université Federal do Paraná. Nous co-dirigeons des étudiant.e.s au Brésil qui font des séjours en France. Nous organisons des journées et des visites de travail avec nos collègues pour avancer sur les thèmes de DOING. Deux séjours ont eu lieu pendant 2022, Carmem Hara a visité le LIFO du 15 août - 10 novembre et Martin Musicante a visité le LIRIS du 7 - 21 novembre 2021.

Actions satellites Le projet APR-IA, financé par la région CVL a été accepté et doit financer deux contrats post-docteraux. Un postdoc dans le domaine du TAL, avec l'objectif de proposer des méthodes (génériques ou issues d'une adaptation au domaine médical) pour la résolution des coréférences et pour l'identification des relations temporelles. Ce travail est porté par le LLL d'Orléans, mais compte avec la collaboration du LIFAT (Tours) et du LISN (Paris-Saclay). Un autre postdoc dont le travail visera le développement d'une première version du système d'interrogation sur des bases de données graphes, dont le langage de requête (déclaratif) engloberait une analyse prédictive. Ce travail est porté par le LIFO d'Orléans, mais compte avec la collaboration du LIFAT (Tours) et du LIRIS (Lyon).

Le projet APR-IA s'insère dans les activités de DOING et nous donne l'occasion d'étendre nos collaborations en région. Ainsi, le groupe de travail TAL se voit renforcé par des collaborations avec des collègues linguistes du laboratoire LLL et les échanges BD-IA se voient enrichis par l'implication de nouveaux chercheurs et de nouvelles chercheures dans le domaine de l'IA au LIFAT. La tâche 2 du projet est entièrement conçue sur cette nouvelle collaboration.

Un *Workshop* associé En septembre 2022, le *workshop* DOING@ADBIS[2], en connexion avec notre action DOING@MADICS, a eu lieu comme événement satellite de la conférence ADBIS[3].

Proposé et organisé par Carmem S. Hara de l'*Universidade Federal do Paraná* (UFPR), Brésil et Mirian Halfeld Ferrari de l'Université d'Orléans, le workshop a reçu 11 soumissions, 4 articles acceptés comme *full papers* et 2 comme *short paper*, ayant ainsi un taux d'acceptation de 50%. Chaque article a été relu par 3 membres du comité de programme. Le comité de programme était composé de spécialistes de différents pays et continents, travaillant dans un des trois domaines phares de DOING : traitement automatique des langues, bases de données et intelligence artificielle.

Le *workshop* a eu lieu en mode hybrid le 5 septembre 2022 au Politecnico di Torino en Italy, avec environ 15 participants.

[2] `https://www.univ-orleans.fr/lifo/evenements/doing/?page_id=551`
[3] `https://adbis2022.polito.it`

Les chairs de DOING@ADBIS comptent re-postuler pour un *workshop* en 2023 au sein de la même conférence.

3 Stages et Groupes de travail

Dans le cadre des appels régionaux, nous avons postulé pour le financement de stages. Nos propositions de stage avaient comme but de continuer des collaborations inter-thématiques débutées l'année dernière. Les stages ont ainsi permis de à nos laboratoires de travaillller en collaboration sur des sujets pratiques et d'avoir des expériences concrètes sur les objectifs et les problèmes abordés dans le cadre de DOING, en général. Avec les stages, nous pensons aussi contribuer à la formation d'étudiants par le recherche, en espérant, principalement pour les plus jeunes, les éveiller à nos thématiques.

3.1 Stage Master : Mise à jour des graphes de propriétés

Il s'agit d'un étudiant financé par un contrat CIFRE. Le but du stage était d'étudier la possible adaptation de la politique de mise à jour des graphes d'attributs, proposée par un groupe de travail au LIFO dans un cadre plus formel, par l'utilisation de Neo4J pour l'implementation de certaines étapes.

4 Bilan scientifique (2022) et perspectives

Le progrès de DOING en 2022 se fait sentir dans les directions de travail et les résultats que nous décrivons dans la suite. Ces résultats sont la conséquence de notre politique de groupes de travail et co-encadrements de stages, permettant une plus grande interaction entre les membres DOING.

Extraction d'information à partir des textes. De nouveaux algorithmes pour l'extraction d'information à partir des textes du corpus DEFT ont été proposés, notamment pour l'extraction de cinq types de relations.

Interrogation et analyse de graphes à partir de requêtes en langue naturelle. Dans les domaines de la recherche d'information, des bases de données et dernièrement de la data science, le traitement des requêtes en langue naturelle est un problème d'actualité qui vise à rendre accessible l'interrogation, l'exploration et l'analyse de données à des personnes avec peu de connaissances en langages d'interrogation formels. Afin de rendre l'interrogation et l'analyse de données et en particulier de graphes aux expert.e.s métier, nous avons travaillé sur la proposition d'une méthode semi-automatique pour traiter les questions de recherche et les transformer vers des requêtes classiques ou data science exécutables sur un gestionnaire de bases de donnée graphes. La caractéristique interactive et exploratoire des requêtes data science appelle à proposer des stratégies de spécification adaptées aux phases différentes de l'expression/évaluation

de ce type de requêtes. Nous allons étudier ces aspects qui nous paraissent importants lorsqu'il s'agit de proposer ce type de modalité d'interrogation à des expert.e.s d'autres domaines avec peu de connaissances en data science.

Vers une cartographie des travaux et des équipes sur la construction et l'interrogation analytique de bases de données graphes. Nous avons fait un premier exercise de cartrographie des thèmes abordés par DOING. Notre champ d'action se trouve au point de croissement de plusieurs façons d'aborder "les graphes" :

- Étude formelle des graphes (i.e., modèles, propriétés, opérations associées);
- Bases de données graphes avec des problématiques de modélisation de bases de données, stockage, indexation et interrogation (les langages et l'exécution des requêtes);
- L'intelligence artificielle qui utilise des modèles de graphes comme outils de représentaiton des connaissances et les modèles d'analyse de graphes issues de l'apprentissage automatique;
- Enfin les domaines de recherche d'information et des traitements de texte basés sur des graphes.

Dans notre cartographie, nous avons choisi comme fil conducteur et comme périmètre d'étude le traitement de textes avec des graphes (représentation de connaissances, bases de données et interrogation). Nous allons poursuivre cette tâche en espérant pouvoir concevoir un article capable d'introduire nos problèmes et un panorama des groupes qui travaillent dessus. Il s'agit d'une tâche laborieuse en cours.

Perspectives Nos avancées montrent que plusieurs défis, intrinsèques au traitement de la langue naturelle, restent d'actualité. En outre, un défi majeur concerne le mapping entre les informations extraites et un schéma de base de données utile à l'utilisateur final.

Nous pensons focaliser nos efforts sur l'analyse des graphes avec des algorithmes IA - pour combiner la gestion des données, l'apprentissage automatique et l'optimisation. Le projet APR-IA porté par M. Halfeld Ferrari Alves avec des participant.e.s de DOING aidera à réaliser cette tâche de manière approfondie.

Activités à venir : éléments de prospective Nous avons travaillé deux ans en tant qu'action et nous allons proposer un projet ajusté selon les expériences observées et les résultats obtenus pendant le deux premières années.

1. Pour la partie extraction d'information, nous faisons face aux défis intrinsèques au traitement de la langue naturelle : traitement de la temporalité, coréférence, distinction entre des actions prescrites et celles effectivement exécutées, etc. Nous avons acquis une vision de la problèmatique

et des verrous, avec un retour important des visions bases de données, analyse de données, TAL, et IA. Nous pourrons orienter la reflexion sur les problèmatiques identifiés et les possibles pistes de recherche de façon plus claire et multi-disciplinaire.

2. Dans la construction d'une instance d'une base de données à partir des textes, un défi majeur concerne le mapping entre les informations extraites et un schéma de base de données utile à l'utilisateur final. La construction d'une instance de base de données à partir d'un texte exige des décisions complexes méritant une analyse assez globale, tant les liens sémantiques sont disséminés dans le texte. La définition de la bonne granularité que la base doit représentée est intimement liée à l'usage envisagé en aval.

3. La construction d'un réseau de collaborateurs spécialistes dans le domaine de la santé (étape que nous avons entamée comme expliqué précédemment). Nous avons démarré cette action en début d'année. Suite à un premier échange avec un certain nombre de professionnelles de santé, nous avons plutôt travaillé à l'ajustement des questions posées dans notre enquête. Une interaction plus approfondie sur les données et leur utilisation dans la pratique de la médicine fera partie du programme de travail de la proposition de reconduction de DOING que nous préparons.

   Via ce réseau, fondamentalement international, nous comptons caractériser les attentes de l'utilisateur final. Cette enquête auprès des utilisateurs cibles est importante pour, d'un coté, définir l'importance de chaque type d'information à stocker et, de l'autre, comprendre les analyses de données à privilégier.

4. Dans le contexte des requêtes *data science*, nous envisageons, en 2023, de nous focaliser sur la conception de requêtes avec des algorithmes IA d'analyse de graphes. Nous espérons combiner la gestion des données, l'apprentissage automatique et l'optimisation. Cela est la base de notre nouvelle collaboration avec le LIFAT.

5. Il sera très important d'établir une typologie de cas d'usage, pour aller vers la construction de solutions plus génériques. Ici, la question consistera à proposer un noyau commun à un type de cas d'usage, pour proposer des solutions à deux couches : un noyau et une couche de personnalisation.

5 Communication DOING

- Site DOING@MADICS : `https://www.madics.fr/ateliers/doing/`
- Site DOING@DIAMS : `https://www.univ-orleans.fr/lifo/evenements/doing/`
- Twitter : `https://twitter.com/NetworkDoing`

- Canal Slack : `https://join.slack.com/t/doing-madics/shared_invite/zt-eg5yg240-pqu~Q2igaDOXs1OpmzOOHg`

6 Bilan financier

Les montants donnés ici sont approximatifs car certaines dépenses sont en cours.

- Missions (2578,26 euros) : prise en charge des missions de cinq personnes pour assister au Symposium MADICS Paris - Lyon, Orléans - Lyon.
- Visites courtes inter-laboratoires : pour la semaine de travail DOING organisée du 7 - 10 novembre, nous avons pris en charge la mission de G. Vargas-Solar et d'un chercheur invité M. Musicante (UFRN, Brésil). *Les repas des collègues d'Orléans n'a pas pu être pris en charge par DOING.*
- Livres (2491,12 euros)

7 Publications associées à DOING

1. Agata Savary, Alena Silvanovich, Anne-Lyse Minard, Nicolas Hiot, Mirian Halfeld Ferrari Alves: Relation Extraction from Clinical Cases for a Knowledge Graph. ADBIS (Short Papers) 2022: 353-365
2. Genoveva Vargas-Solar, Karim Dao, Mirian Halfeld Ferrari Alves: NLDS-QL: From natural language data science questions to queries on graphs: analysing patients conditions & treatments. CoRR abs/2208.10415 (2022)
3. Genoveva Vargas-Solar, Pierre Marrec, Mirian Halfeld Ferrari Alves: Comparing Graph Data Science Libraries for Querying and Analysing Datasets: Towards Data Science Queries on Graphs. ICSOC Workshops 2021: 205-216