

Données Intelligentes : transformer l'information en connaissance Proposition ACTION MADICS

Mirian Halfeld Ferrari Alves, LIFO, Orléans

Anne-Lyse Minard-Frost, LLL, Orléans

Genoveva Vargas-Solar, LIRIS, Lyon

Acronyme : DOING

Mots-clés : extraction d'information, graphe de propriété, base de données graphe, conception de bases données guidée par des requêtes, langage de requêtes et mises à jours sur graphe (GQL, Open Cypher), interrogation en fonction du contexte, contraintes d'intégrité, qualité, intelligence artificielle, data science.

Thématiques concernées : bases de données, traitement automatique du langage naturel, intelligence artificielle (techniques pour le traitement de texte et *data analytics* sur les textes/documents).

Types de données considérées : Notre intérêt porte sur tout type de données textuelles non structurées (ou éventuellement semi-structurées) à partir desquelles nous voulons construire des bases de données qui puissent être explorées, interrogées et analysées pour extraire des connaissances. Parmi les domaines d'application, jusqu'à maintenant, nous avons travaillé surtout sur les données du domaine médical. Néanmoins DOING reste intéressé non seulement aux données textuelles du domaine de la santé mais aussi du domaine de l'environnement, et éventuellement aux textes juridiques concernant ces deux domaines. Nous considérons des textes provenant de différents types de documents (articles scientifiques, forums, réseaux sociaux, rapports d'activité, formulaires semi-structurés, etc.). Un aspect plus original de DOING concerne l'utilisation de collections de questions formulées par des spécialistes du domaine d'application pour calibrer l'extraction d'information et de connaissance.

Nous avons plusieurs jeux de données à disposition que nous pourrions utiliser dans le cadre de cette action, à la fois en français et en anglais. Pour le **domaine médical** nous pouvons par exemple citer :

1. Corpus de cas cliniques développé et annoté pour DEFT 2019/2020 (<https://deft.limsi.fr/>) et accessible après acceptation de la licence ;
2. Corpus de cas cliniques en cours de développement (fin du projet en juillet 2021) multilingue (français, anglais, italien, espagnol, basque) dans le cadre du projet ELG (European Language Grid) E3C (<https://e3c.fbk.eu/home>); licence CC-BY-NC-SA ;
3. QUAERO FrenchMedical corpus (https://quaerofrenchmed.limsi.fr/quaeroFRmed_FR.htm) qui contient entre autres des notices de médicament annotées ;
4. COVID-19 Open Research Dataset (non annoté) : <https://allenai.org/data/cord-19>
5. S2ORC: The Semantic Scholar Open Research Corpus (non annoté) : <https://github.com/allenai/s2orc>

Pour le **domaine de l'environnement**, nous avons pour initier des recherches par exemple :

- S2ORC: The Semantic Scholar Open Research Corpus (non annoté) : <https://github.com/allenai/s2orc>
- PANACEA Environment French monolingual corpus (non annoté) : <https://catalogue.elra.info/en-us/repository/browse/ELRA-W0065/>
- Nous signalons les interactions possibles avec le programme régional ARD-CVL-JUNON (voir Bilan section 5). Dans ce cadre, seulement le volet '*interrogation intelligente*' de notre action serait concernée, en utilisant des données (non textuelles) que JUNON compte laisser disponibles.

Domaines d'applications scientifiques : santé publique (stratégies thérapeutiques, pharmacovigilance, toxicologie), et environnement continental (transition écologique, transport, gestion de l'eau, agriculture, villes intelligentes)

Enjeux sociétaux : Les enjeux sont liés aux possibles améliorations dans le domaine de la santé publique et l'environnement suite à la capacité d'analyse, d'intégration, d'interrogation intelligente et conviviale des bases de connaissances construites à partir des textes du domaine. La connaissance acquise à travers l'évaluation de requêtes d'exploration, d'interrogation et d'analyse de données pourra non seulement être mise à disposition des professionnels des métiers de la santé et de l'environnement, mais aussi d'un individu isolé (par exemple, un patient ou un citoyen).

Dans le domaine de la santé, les enjeux sociétaux sont nombreux : détection des problèmes liés aux médicaments ; mise en relation des caractéristiques physiologiques des patients et leurs pathologies ; demande de traitement individualisé ; optimisation des parcours médico-économique pour certaines pathologies ; évaluation comparative de l'efficacité de stratégies thérapeutiques : impact des changements de l'environnement sur la santé publique ...

Dans le domaine de l'environnement, l'objectif est d'interroger les données collectées pour extraire de la connaissance et ainsi contribuer à la solution de problèmes liés à la performance énergétique, à l'écologie et à l'environnement qui sont soumis aux changements globaux.

Enjeux économiques : La science des données est une réponse à la problématique de la gestion des politiques de santé publique et de l'environnement. Les budgets gouvernementaux attribués aux dépenses de santé et de l'environnement sont souvent très conséquents. De grandes sociétés mondiales investissent également dans ces domaines car ce sont des enjeux majeurs pour assurer la santé des humains dans un environnement propre. L'optimisation des parcours médico-économique, la définition de stratégies et politiques publiques et les avancées scientifiques et technologiques sont au cœur des solutions envisageables pour des économies de la santé et de l'environnement satisfaisantes et efficaces.

Résumé de l'Action

L'action DOING se concentre sur la transformation des données en information puis en connaissance. L'idée est de mettre à contribution les expertises des chercheurs en TAL, BD et IA pour :

(1) extraire des informations dans les données textuelles et les représenter pour peupler des bases de connaissances. Le choix des entités et des relations à extraire pourra être guidé par les requêtes et éventuellement la structure de la base, avec l'idée d'extraire et stocker les données les plus intéressantes et importantes pour les consultations et les analyses des experts. DOING s'intéresse en particulier à la question de l'extraction de relations (pas uniquement binaire et pas uniquement interprétable) et à ce que l'interaction avec les bases peut apporter à cette tâche ;

(2) proposer des méthodes intelligentes pour la manipulation et la maintenance de ces bases avec de nouvelles formes de requêtes englobant des mécanismes d'analyse efficaces, flexibles, sûrs, adaptés à l'utilisateur et qui respectent des contraintes de qualité et de vie privée. DOING s'intéresse en particulier aux graphes de propriétés, au langage d'interrogation style Cypher dans le cadre de Open Cypher et GQL et aux algorithmes d'analyse des graphes (centrality, community detection, similarity, link prediction, pathfinding).

Contexte scientifique de l'Action *

Les textes sont une source d'information et de communication importante pour les métiers de la santé. Il en existe différents types : protocoles médicaux, articles scientifiques, dossiers des patients, rapports d'intervention, suivis des traitements, tweets, observations des l'environnement, etc. Dans le domaine de l'environnement, beaucoup d'informations circulent aussi entre les experts sous forme textuelle, comme par exemple dans les rapports des bureaux d'études, les articles scientifiques, ou encore les textes législatifs.

L'exploitation intelligente, efficace et sûre des documents requiert, dans la plupart des cas, la transformation de l'information non structurée présente dans ces textes, dans un ensemble de données, structurées selon une sémantique adaptée au domaine d'application. Plus encore, cette étape de structuration devrait être faite en visant à écarter les superflus, pour se concentrer sur les données essentielles aux applications et analyses en aval. En effet, la résolution des problèmes de la société moderne passe par une recherche interdisciplinaire et les croisements des informations, surpassant ainsi une simple mise à disposition de données non structurées. Il existe un besoin urgent d'une représentation plus souple, plus fine, plus sensible au contexte, facilement accessible via des outils de consultation. En outre, la consultation, dans ce cadre, demande des requêtes robustes, efficaces, proposant une analyse des informations, avec une assurance de qualité, de cohérence et du respect de la vie privée. Les mêmes principes doivent régir la maintenance et l'évolution des bases de connaissances. Un tel environnement représente la nouvelle génération des infrastructures, la mise en place de la science de données au bénéfice du progrès scientifique ou de la prise de décisions éclairées.

Cette action s'intéresse à toutes les étapes concernant la mise en place de ces infrastructures modernes, se focalisant sur les données liées à la santé et à

l'environnement. Elle vise en particulier deux grandes lignes de discussion ainsi que leur mise en relation :

(I) **La transformation des données en information**, c'est-à-dire, l'extraction de l'information des données textuelles pour peupler une base de connaissance. Cette extraction peut englober des questions diverses comme la détection des entités, leur normalisation et l'extraction des relations entre entités ; la construction d'un schéma à partir des instances ; les *mappings* entre les schémas ; le résumé des données ; les problèmes d'anonymisation des données à publier ; l'extraction de contraintes d'intégrité à partir des textes juridiques ; etc. Après un premier travail sur l'extraction des entités cette année (voir Bilan, section 3.1), nous souhaitons dans l'action DOING mettre l'accent sur l'extraction des relations, et sur l'interaction des systèmes d'extraction d'information et des bases de données (du texte à la base de données et de la base de données au texte). Nous voulons aussi, en lien avec ces aspects, réfléchir aux problématiques de l'adaptation au domaine et de la fusion d'information provenant d'articles scientifiques, de cas cliniques, de base de données des médicaments, etc.

(II) **La transformation de l'information en connaissance**, c'est-à-dire, l'interrogation intelligente et efficace, et la maintenance de bases de connaissances. L'intelligence dans la manipulation d'une base de connaissance passe par la capacité d'offrir des mécanismes d'analyse efficaces, flexibles, faciles à utiliser et adaptés à l'utilisateur. Les requêtes '*data science*', capables de rendre des résultats analytiques sur les données de la base, sont des perspectives importantes dans l'aide à la décision et devraient être considérées en fonction de l'efficacité, la fiabilité, le profil de l'utilisateur, le respect des contraintes, du contexte, de la vie privée... Suite à une première année de travail en tant qu'atelier (voir Bilan, section 6), DOING propose une réflexion plus centrée sur les graphes de propriétés, en se focalisant, dans un premier temps, sur les valeurs classiques du domaine bases de données (en contraste avec le domaine du web sémantique), en particulier en ce qui concerne la dichotomie entre les sémantiques du monde ouvert et fermé. DOING s'intéresse en particulier au langage d'interrogation style Cypher dans le cadre de Open Cypher et GQL et aux algorithmes d'analyse des graphes (centrality, community detection, similarity, link prediction, pathfinding).

Positionnement dans le paysage de la recherche *

L'atelier DOING réunit des chercheurs en traitement automatique du langage naturel, en bases de données et en intelligence artificielle autour de la problématique des données intelligentes.

Initialement proposé comme groupe de travail du réseau régional RTR-DIAMS, DOING travaille au niveau régional depuis 2019 avec les laboratoires LIFO, LIFAT, et LLL. Dès cette période, DOING a commencé à s'étendre vers des collaborations nationales et internationales. Son évolution vers un atelier MADICS a naturellement accentué cette expansion. Aujourd'hui nous comptons avec plusieurs groupes ayant manifesté leur intérêt pour nos activités (voir liste ci-dessous) et nous avons 40 déclarations d'intention sur le site MADICS de DOING. Nous croyons que cette première année d'activité de DOING@MADICS, malgré la crise sanitaire, a permis de consolider ou de créer des liens entre, non seulement les laboratoires de la région

Centre-Val-de-Loire, le LIFO, le LIFAT, le LLL, mais aussi de ces laboratoires vers le LIPADE, l'IRISA (qui participe à nos réunions depuis 2019), le LIG, et d'autres encore, comme nous pouvons constater via les déclarations de soutien (ci-dessous) et les échanges réalisés en 2020 (voir Bilan). Nos liens internationaux ont aussi été renforcés avec le Brésil : avec des anciens partenaires, UFRN et UFPR, ce dernier co-porteur du workshop international DOING@ADBIS-TPDL-EDA et le démarrage de nouvelles collaborations, notamment avec des médecins de l'hôpital universitaire de l'UFJF. Un contact prometteur avec l'Université de Genève a été aussi établi. De plus, il ne faut pas oublier que l'atelier compte avec la participation de chercheurs de l'UMI LAFMIA au Mexique. Enfin, une thèse Cifre-LIFO [1], dans la thématique de DOING, formalise une collaboration avec une entreprise capable d'exprimer des besoins pratiques dans la construction des logiciels pour le domaine de la pharmacovigilance.

Avec l'objectif d'élargir le réseau et de le consolider à travers la participation d'autres groupes intéressés à la définition des nouvelles techniques d'interrogation de données et d'extraction de connaissances, nous voulons créer cette action MADICS. Des groupes experts en extraction d'information, en représentation de la connaissance, en data analytics et en traitement de requêtes pourront réfléchir sur des problèmes communs et sur des perspectives ouvertes.

Pour le positionnement nous avons trouvé intéressant de voir ce que nous avons en commun avec certains Ateliers/Actions existants et ce qui est particulier à DOING.

1. Nous avons remarqué que MADONA, HN, LEMON et DOING partent du traitement de données textuelles, souvent en utilisant des techniques de TAL mais bien sûr avec des objectifs différents. RoD, LEMON et DOING ont en commun de se servir des graphes pour représenter le contenu des textes, mais à la différence de RoD qui ne travaille pas forcément uniquement avec du texte, LEMON et DOING se concentrent sur les collections textes.
2. LEMON et RoD s'intéressent en particulier aux ontologies. RoD s'intéresse à l'intégration de données par des ontologies alors que LEMON, DOING et MADONA ne travaillent pas sur l'intégration de données mais plus sur une phase de « préparation de données » qui leur permet d'extraire du contenu et de le représenter pour ensuite l'interroger, appliquer de la fouille, ou du ML pour en extraire de la connaissance.
3. DOING a la particularité lors de cette phase de préparation de données de s'intéresser à la qualité exprimée par des contraintes, et pour l'interrogation d'étudier comment exprimer de manière déclarative les tâches de traitement des graphes et comment les optimiser comme dans le cas des requêtes relationnelles.
4. Certaines questions abordées dans l'item (II) de notre contexte scientifique, sont aussi sujets de réflexion dans l'action RoD. Néanmoins, contrairement à RoD, l'action DOING a des préoccupations plus accrochées au domaine des bases de données - et pas forcément ancrée dans le domaine du web sémantique. Cette différenciation, subtile parfois, nous guide vers des choix différents dans plusieurs situations. DOING propose de donner priorité aux choix issus du domaine 'base de données' et d'organiser des échanges avec les actions travaillant sur le web sémantique, car si nous sommes persuadées que les liens existent et que l'échange serait riche, nous sommes aussi convaincues que le travail en parallèle ne peut qu'apporter des évolutions scientifiques.

5. L'action ATLAS (Apprentissage et optimisation Large-échelle : Application à la Santé). L'ambition de DOING est de ne pas se restreindre à un domaine d'application, mais le domaine de la santé est une de ses priorités. Contrairement à ATLAS qui considère des données hétérogènes comme les images ou les données provenant des capteurs, DOING se concentre sur l'extraction d'information des textes. En outre, ATLAS est une action centrée sur les questions d'IA, en particulier les techniques d'apprentissage automatique. Le point de vue de DOING est centré bases de données : construction, manipulation, évolution, interrogation. DOING s'intéresse aux algorithmes d'apprentissage automatique comme éléments à intégrer dans un langage de requête.
6. L'atelier MADICS-HN questionne les données comme objet de leur recherche et ce que les domaines des SHS et de l'informatique ont à apporter l'un à l'autre. Leur recherche porte également sur des données multimédia contrairement à nous. Dans ce cadre, la question de l'hétérogénéité et le rôle des métadonnées sont des aspects que MADICS-HN vise à examiner avec un regard multidisciplinaire. Leurs objectifs, bien que d'intérêt aussi pour nous, ne constituent pas les activités centrales de DOING.

Trois responsables (Justificatif). Comme déjà indiqué auparavant, DOING@MADICS a son origine dans un groupe de travail régional porté par le LIFO. La proposition d'atelier faite à MADICS comptait déjà trois responsables issues des laboratoires suivants : le LIG, comme membre INS2I, le LLL comme membre CNRS non-informaticien et le LIFO comme membre fondateur de l'initiative. Il faut ici souligner que le partenariat LIG-LLL-LIFO a commencé dans DOING@DIAMS. Aujourd'hui, après une année de travail que nous considérons riche et agréable, la proposition de DOING comme action compte trois responsables (les mêmes que pour l'atelier) avec un petit changement dû à la récente affectation de Genoveva Vargas-Solar au LIRIS. Le travail à trois se justifie aussi en terme scientifique avec Anne-Lyse Minard-Forst (LLL) épaulant le côté traitement automatique des langues, Genoveva Vargas-Solar (LIRIS) sur les aspects d'interrogation efficace des bases de données et en particulier des requêtes data science, et Mirian Halfeld-Ferrari (LIFO) sur les aspects plus théoriques des bases de données en vue des liaisons avec l'extraction d'information et l'extraction de connaissances. Notre collaboration est étroite et les différentes actions menées au cours de l'année 2020 montrent notre engagement à motiver la mise en place des échanges multidisciplinaires, dans les domaines de DOING. Par nos points de vue, différents, et nos tempéraments divers et complémentaires, il nous semble que nous avons toutes contribué à faire progresser DOING. Nous souhaitons continuer dans cette direction.

Objectifs de l'Action et principales activités envisagées *

Nous prévoyons les activités suivantes :

- a) Journées d'étude permettant aux acteurs du domaine :
 - d'échanger leurs points de vue sur le problème global;
 - de répertorier les besoins de la communauté ou des entreprises;
 - d'ouvrir un espace aux doctorants travaillant sur le domaine - avec une présentation de leurs travaux ou projets de recherche.

b) Organisation de réunions récurrentes en sous-groupes d'intérêt pour travailler sur des problèmes liés à la construction de bases de données/connaissances à partir de l'analyse des textes et de leur exploration.

c) Organisation de tutoriaux pour acquérir une connaissance pratique des outils et infrastructures pour le traitement des textes, la construction, l'interrogation et l'analyse de graphes. Une première suite de tutoriaux a été proposée dans le cadre de l'atelier DOING.

d) Définition et encadrement de stages, sujets de master et de doctorat conjoints entre partenaires pour avancer sur les questions étudiées dans le cadre de DOING.

e) La préparation d'un article de synthèse sur la thématique de la transformation des données en information, puis en connaissance. Cet article devra être soumis à une revue.

f) Deuxième/troisième éditions d'un workshop sur la thématique (coopération LIFO-UFPR dans le cadre la conférence internationale ADBIS).

g) Co-organisation de journées communes entre ateliers/actions pour aborder des problèmes/thèmes communs à leurs projets. L'objectif sera d'apporter différentes perspectives et discuter différentes stratégies et contextes dans lesquelles ces problèmes émergent et comment ils se déclinent.

Les résultats attendus de l'action DOING-MADICS ainsi que les indicateurs de son succès sont :

- La réussite de toutes ses activités (en particulier l'article de synthèse, le workshop et les co-encadrements).
- Un échange plus large, au niveau national, autour de la question de la transformation des données en connaissance, englobant les questions qui vont du traitement automatique du texte à l'interrogation 'data science'.
- Une plus grande visibilité à travers la déclaration d'intérêt, nombre de participants lors des activités, rédaction de papiers, use cases, proposition de projets avec des partenariats issues de DOING.
- Une augmentation (quantitative et qualitative) de la production scientifique, grâce à l'accroissement de l'échange scientifique.

Fonctionnement envisagé *

Planning prévisionnel

Année 1:

- Webinar: Le kick-off de DOING: discussions autour d'un exposé invité, permettant d'entamer la dynamique du groupe.
- Journée de travail 1:
 - échange sur le domaine de manière global (keynote, présentations courtes de différents travaux avec une partie consacrée aux doctorants);

- discussions sur la formation de sous-groupes thématiques de travail, tentative de composition des sous-groupes. En préparation pour une nouvelle journée d'étude, chaque sous-groupe devrait avoir comme un de ses objectifs de préparer une présentation 'grand public' à l'assemblée DOING (problèmes, directions de recherche envisagées en priorité, etc);
- table ronde sur l'organisation de l'**article de synthèse**.
- Journée de travail 2 :
 - présentation des sous-groupes thématiques de travail
 - définition de la structure de l'article de synthèse, les contributions à préparer et les dates limites
- Workshop -- 2nd Workshop DOING@ADBIS (candidature en préparation)

Année 2:

- Journée de travail 3 :
 - discussion sur des points précis (définis en amont);
 - présentation des sous-groupes de travail thématiques avec, éventuellement, présentation des travaux, stages, projets de thèse dans le domaine;
 - discussions autour de l'article de synthèse
- Journée de travail 4 :
 - bilan et perspectives de l'action,
 - définition d'un projet scientifique en vue d'une soumission à un appel à projet.
- Finalisation et soumission de l'article de synthèse.
- Workshop -- 3rd Workshop DOING@ADBIS (re-candidature si possible)

L'aide financière demandée concerne :

- l'organisation des journées de travail et
- l'invitation de personnalités extérieures à l'action aux journées de travail/workshop.

Planning global de l'aide financière (sur un total de 3000 euros pour deux ans) :

- Année 1 : 1500 euros
- Année 2 : 1500 euros