# DOING@MADICS

Bilan 2020 (détails)

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

November 2020

## 1 Introduction

DOING a été accepté comme un atelier MADICS en 2020, année durant laquelle nous avons lancé une série d'actions pour favoriser des collaborations. Notre atelier 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 atelier :

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

La réflexion et le travail sur ces enjeux sont faits avec une perspective multi-disciplinaire avec notamment les domaines du *traitement automatique des langues, des bases de données et de l'intelligence artificielle*. Cette année, notre réflexion a été associée plutôt au domaine d'application de la santé, mais le domaine de l'environnement reste un de nos objectifs.

DOING@MADICS représente aussi une extension nationale du groupe de travail régional DOING@DIAMS qui motive des collaborations au sein de la Région Centre Val du Loire.

## 2 Organisation d'événements

L'année 2020 a été marqué par la crise sanitaire hors norme. Notre planning d'activités a ainsi été impacté, comme beaucoup d'autres. Nous l'avons adapté aux circonstances. Dans cette section, nous proposons un bilan organisé par type d'activité ; les activités sont décrites dans un ordre chronologique.

## 2.1 Journées d'études et *webinars*

Les activités d'animation de DOING ont commencé par l'organisation d'une journée d'étude[1] qui devait avoir lieu le 8 avril, à l' INALCO (65 rue des Grands Moulins 75214 Paris Cedex 13 – SALLE 3.13). Cet événement a malheureusement été ajourné, en raison du confinement. La journée en "présentielle" n'a pas pu être reprogrammée depuis. Nous rappelons ci-dessous le planning prévisionnel de cette journée qui incluait également une présentation des objectifs DOING et des séances de discussions et synthèse sur les thématiques abordées pendant les exposés.

- **Deux exposés longs** : nos invitations avaient déjà été acceptées par Vasiliki Foufi, Université de Genève et Salima Benbernou, Université Paris Descartes, LIPADE.
- **Quatre présentations courtes** : un appel à proposition (extrait ci-dessous) avait déjà été lancé dans les listes de diffusion MADICS et autres, suite auquel, à la date de l'annulation, quatre propositions avaient déjà été reçues.

  > L'atelier DOING invite les chercheurs de la communauté MADICS à participer à une journée d'études le 8 avril 2020 pour mener une réflexion sur les données intelligentes du point de vue des bases de données, de l'intelligence artificielle, du traitement automatique du langage naturel et le traitement de données. [...]
  >
  > Nous invitons également les chercheurs qui souhaitent assister à cette journée à présenter leurs travaux (15 min + 5 minutes de questions) sur leurs thèmes de recherche en cours ou déjà publiés associés aux deux grandes lignes abordées par DOING et leur mise en relation [...]
  >
  > Ceux qui souhaitent proposer une présentation doivent nous envoyer à l'adresse doing.madics@gmail.com avant le 25 mars 2020 : a) Titre de la présentation, b) Auteurs : nom, institution, email et c) Résumé de votre présentation (10 – 15 lignes)[...]

- **Séance poster pour des doctorants** : ici aussi, un appel à proposition avait déjà été lancé, l'idée étant d'inviter des étudiants en thèse à proposer et présenter un poster illustrant leurs travaux ou idées en cours dans une des deux grandes lignes abordées par DOING. À la date de l'annulation nous comptions six propositions.
- **Séance de présentation de l'atelier MaDICS-HN** : nous avons suivi le conseil de l'équipe MaDICS et nous avons invité les collègues de l'atelier MaDICS-HN à venir faire une présentation de leurs objectifs dans cette première journée.

Nous avons estimé pouvoir réunir une vingtaine de personnes pour cette journée d'étude (à la date du 9 mars nous avions 10 inscrits).

[1] `https://www.univ-orleans.fr/lifo/evenements/doing/?page_id=259`.

Dans la suite, avec l'annulation de la journée vers le 16 mars, nous avons décidé d'organiser des *webinars*[2] - relativement courts, limités à 2 heures au maximum - pour remplacer la journée d'étude annulée. Le but était de commencer les discussions autour des sujets DOING, en profitant des invitations déjà faites et en proposant des réunions plus courtes, plus adaptées aux conditions de la visio-conférence. Nous avons ainsi organisé deux *webinars*, dont les programmes incluaient un *Keynote* et deux exposés courts. Nous avons aussi essayé de maintenir un équilibre entre les deux enjeux thématiques de DOING.

### 2.1.1 Webinar 1 : 05 Juin 2020

Le programme de ce premier *webinar* était le suivant:

- 10:00 – 11 :00 - **Keynote : Managing data quality in the age of big data.** Prof. Salima Benbernou, Université Paris Descartes, LIPADE.
- 11:00 – 11:25 - **Alignement de bases de données pour l'extraction d'informations concernant les sols pollués.** Chuanming Dong, LASTIG, Univ Gustave Eiffel, ENSG, IGN
- 11:25 – 11:45 - **DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées.** Andreane Roques, Université d'Orléans, LIFO

Le problème de la cohérence d'une base de données face au volume des données (aspect important dans le Keynote de Salima Benbernou) et la conception des bases structurées à partir de données non structurés (thématique abordée par des différents points de vues dans les exposés courts) ont été discutés dans cette première rencontre qui a compté 26 participants (détails dans le tableau 1).

Nous avons organisé une discussion à partir de questions qui ont été soulevées par le keynote sur les contraintes et la qualité de données. Concernant les contraintes de données il s'agit d'identifier le type de contraintes à considérer surtout lorsque des données sont issues de sources différentes. Un aspect clé de la discussion a abordé l'utilisation des contraintes lors de l'interrogation de données.

La préoccupation avec la qualité des données (en assurant leur cohérence) est un point commun entre les travaux de l'équipe de Salima Benbernou et les membres de l'équipe "bases de données" du LIFO. Le webinar DOING a permis un premier contact, avec l'envie d'entamer un sous-groupe de travail au sein de DOING, probablement en proposant un stage co-encadré en 2021. De manière plus large, les aspects concernant la prise en compte des contraintes et la provenance pendant l'interrogation intéresse aussi les membres d'autres laboratoires comme le LIG et le LIRIS. L'exposé de Chuanming Dong abordait le problème de l'alignement de plusieurs bases de données afin de les extraire et

[2] `https://www.univ-orleans.fr/lifo/evenements/doing/?page_id=448`

de les réorganiser en une seule base. Dans son exposé, Chuanming Dong a mentionné l'utilisation des techniques TAL dans le but d'améliorer l'appariement des bases. Même si le travail nous a semblé pertinent et intéressant nous n'avons pas pu constaté des liaisons immédiates avec les participants du webinar. L'exposé d'Andreane Roques concernait le résultat d'un stage (décrit ci-dessous) dont le but était d'entamer la collaboration entre les chercheurs en TAL et BD. Lancé dans le cadre des échanges DOING@DIAMS, ce stage peut être considéré comme très positif et nous planifions de continuer en 2021 le travail qu'il a permis de démarrer.

#### 2.1.2 Webinar 2 : 17 Juin 2020

Le 17 juin le *webinar* DOING a suivi le programme ci-dessous et a compté 36 participants (tableau 1).

- 10:00 – 11:00 - **Keynote : Natural Language Processing in the Health Domain**, Dr. Vasiliki Foufi, University of Geneve
- 11:00 – 11:25 - **On Anonymizing the Provenance of Collection-Based Workflows**, Khalid Belhajjame, PSL, Université Paris-Dauphine, LAMSADE
- 11:25 – 11:45 - **Exploring COVID-19 documents with the S-COVID engine**, Mehrdad Farokhnejad, Université de Grenoble, LIG

Cette fois, alors que le *Keynote* proposait un panorama du traitement automatique du langage naturel dans le domaine de la santé, les exposés courts se sont focalisés sur différents aspects des traitements des données.

Nous avons organisé une discussion à partir de questions qui ont été soulevées par le keynote sur l'utilisation de données médicales, les aspects sécurité, comme, par exemple, l'anonymisation de données pour pouvoir mener des analyses. L'intégrité des données a été aussi un point important surtout en considérant les aspects sécurité et vie privée.

Vasiliki Foufi est chercheuse au Département de radiologie et d'informatique médicale de l'Université de Genève. Titulaire d'un doctorat en linguistique computationnelle, elle s'est intéressée à DOING, et elle est ouverte, à l'avenir, aux possibilités de projets internationaux. Dans son keynote, elle a dressé un bilan des applications du TAL, principalement en extraction d'information, pour l'exploitation de données médicales. Travaillant en étroite collaboration avec des professionnels de la santé et sur des données de l'hôpital, nous pourrons bénéficier de son expertise sur la problématique de l'anonymisation (ou déidentification) ainsi qu'éventuellement de son expérience dans la mise en place de projets avec des experts du domaine médical.

Concernant la provenance de données et leur anonymisation dans le cadre de workflows scientifiques nous trouvons qu'il y a des problèmes communs avec les requêtes *data science*. En particulier, concernant la classification de requêtes qui peuvent être posées pour analyser les données et la traçabilité de leur exécution.

| | Institutions des intervenants | Participants |
|---|---|---|
| **Webinar 1** (05 juin 2020) | | 26 participants |
| **Qualité de données** | U. Paris Descartes, LIPADE | **Labs/Instituts (5)** : LIFAT, ISEP, BRGM, etc. |
| **Alignement des BD** | U. Gustave Eiffel, ENSG, IGN, LASTIG | **UMR (7)** : IRISA, Telecom-Paris Tech, LIRIS, LIPN, etc. |
| **Annotation des entités cliniques** | U. d'Orléans, LIFO | **International** : Eurecat |
| **Webinar 2** (17 juin 2020) | | 36 participants |
| **TAL pour le domaine de la santé** | U. de Genève, Suisse | **Labs/Instituts (9)** : LJL Lagrange, OCA, SESTIM, ISPED, INSERM, ESILV, ENGIE, etc. |
| **Anonymisation et provenance** | U. Paris Dauphine, LAMSADE | **UMR (12)** : IRIT, LAMSADE, TETIS, LRI, LIRMM, etc. |
| **Exploration de documents COVID-19** | U. Grenoble Alpes, LIG | **International** : U. Genève, Suisse ; ULB ; U. Manuba, Tunisie |
| **Webinar 3** (17 juillet 2020) | | 40 participants [3] |
| | U. of Manchester | **Labs/Instituts** : ERIAS Inserm, INRAE, AgroParisTech, etc. |
| | Institut de Recherche pour le Développement (U228) | **UMR** : LAAS, LIMSI, etc. |
| | | **International** : Universidad de las Américas, Puebla ; University of Manouba, ENSI, Tunisia |

Table 1: DOING@MADICS' Webinars

Les workflows scientifiques vus comme des requêtes *data science* peuvent être appliqués pour analyser des collections de données textuelles et en extraire des connaissances. Le travail de Khalid Belhajjame s'approche également des workflows conçus pour explorer des collections de données textuelles traitées par le prototype S_COVID présenté lors du Webinar 2 par Mehrdad Farokhnejad. Une réflexion sur ces aspects a commencé à s'organiser avec K. Belhajjame sur les requêtes *data science* sur le traitement et l'analyse de textes et la provenance.

### 2.1.3 Second Symposium GDR CNRS MADICS : Webinar DOING

Le 17 juillet 2020, dans le cadre du Symposium MADICS, DOING a organisé une demi-journée avec deux *Keynotes* et 40 participants se sont joints au webinar.

- 14:00 – 14:10 - Introduction et présentation de la demi-journée
- 14:10 – 15:00 - **Keynote : From Deep Learning to Deep Semantics** , Dr. Andre Freitas, University of Manchester
- 15:00 – 15:20 - Pause café on-line
- 15:20 – 16:10 - **Keynote : Data Cleaning and Preparation for ML and Data Analytics: Toward a Principled Approach**, Dr. Laure Berti-Équille, Institut de Recherche pour le Développement (U228), France
- 16:10 – 17:00 - Discussion autour de questions soulevées lors des différents webinars DOING

[3] Nous n'avons pas récupéré la liste des inscrits et leurs affiliations. Nous avons noté la présence de 40 participants au maximum au cours de ce Webinar, mais nous n'avons pas la liste complète des laboratoires de ces participants.

Les discussions ont porté sur les problèmes et verrous lors de la construction d'une base de données graphe à partir d'un ensemble de textes, notamment lors de la détection des relations. Nous avons discuté suite à la présentation d'André Freitas sur le processus de généralisation des relations (proches du texte) vers des relations proches d'une base de données. Il s'agit par exemple de réfléchir à la redondance des données pour ne stocker que des données représentatives du contenu des textes.

Le nettoyage des données et leur préparation pour les utiliser dans des processus de data analytics et d'apprentissage automatique (keynote L. Bérti-Equille) nous a poussé à réfléchir à l'expression déclarative des pipelines (i.e., requêtes data science) pour définir des processus de préparation de données. Nous avons identifié des techniques pour préparer les données qui sont proches à celles trouvées pour la construction de bases de données à partir du traitement des textes. Les travaux d'André Freitas et de Laure Bérti-Equille proposent deux façons différentes mais complémentaires pour traiter les textes et en construire des bases de données utiles pour l'interrogation, la récupération d'information et d'autres processus d'analytics qui utilisent des modèles d'IA. Cette différence de perspective montre l'ouverture et la richesse des thèmes abordés dans DOING.

## 2.2 Un *Workshop* associé

En août 2020, le *workshop* DOING@ADBIS-TPDL-EDA[4], en connexion avec notre atelier DOING@MADICS, a eu lieu comme événement satellite de la conférence ADBIS-TPDL-EDA[5].

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 17 soumissions, 8 articles acceptés comme *full papers* et 1 comme *short paper*, ayant ainsi un taux d'acceptation de 50%. Chaque article a été relu par 3 membres du comité de programme. Pour une première édition d'un *workshop*, ce résultat a été considéré comme un succès, et la bonne qualité de la plupart des articles soumis a été reconnue. 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 par visio-conférence le 25 août 2020, avec environ 35 participants. Les communications acceptées ont été réparties dans deux sessions techniques du programme : *traitement automatique des langues pour l'extraction de l'information* (avec 4 articles) et *gestion intelligente des données* (avec 5 articles). Le programme de l'atelier comportait également une conférence invitée : Marie-Christine Rousset, professeur et membre du Laboratoire d'Informatique de Grenoble (LIG).

Les chairs de DOING@ADBIS-TPDL-EDA comptent re-postuler pour un *workshop* en 2021 au sein de la même conférence. Nous souhaiterions rendre pérenne cette rencontre.

[4] `https://www.univ-orleans.fr/lifo/evenements/doing/?page_id=77`
[5] `http://eric.univ-lyon2.fr/adbis-tpdl-eda-2020/satellite-events/workshops/`

# 3 Groupes de travail et stages

Dans le cadre des appels régionaux liés à DOING@DIAMS, nous avons eu l'occasion de postuler pour le financement des stages. Les propositions faites avaient comme but d'entamer des collaborations inter-thématiques. Les stages ont ainsi permis de coopérer à travers des sujets pratiques entre nos laboratoires et d'avoir des expériences concrètes sur les objectifs et les problèmes abordés dans le cadre de DOING, en général.

## 3.1 Stage : Extraction de relations

Nous avons obtenu un financement par la fédération ICVL[6] (Informatique Centre Val de Loire) pour un stage de 6 mois sur la thématique de l'extraction de relations entre entités dans le domaine médical. Le stage a été encadré par Mirian Halfeld Ferrari (LIFO, Orléans), Anne-Lyse Minard (LLL, Orléans) et Agata Savary (LIFAT, Tours), toutes trois membres de DOING (@MADICS et @DIAMS). Nous avons recruté une étudiante de M2 (master sciences du langage, parcours Linguistique Outillée et Traitement Automatique des Langues, de l'université d'Orléans), Andréane Roques. Le stage a conduit à la réalisation (i) d'un état de l'art sur l'extraction d'entités et l'extraction de relations dans le domaine médical ; (ii) au développement d'un système pour l'annotation automatique d'entités cliniques (participation à la tâche DEFT[7]) ; (iii) à un travail exploratoire sur l'extraction de relations.

## 3.2 Stage : Data Science Pipelines sur les graphes

Ce stage a été dirigé par Genoveva Vargas-Solar (LIG, Grenoble), Bich Dao (LIFO, Orléans), Mirian Halfeld Ferrari (LIFO, Orléans), et Christel Vrain (LIFO, Orléans) membres de DOING (@MADICS et @DIAMS). Le travail a été réalisé par Pierre Marrec, étudiant-normalien en L3 à l'ENS de Lyon, avec un financement du laboratoire LIFO. Il a contribué à la conception de nouvelles formes pour traiter et interroger les données en combinant des techniques d'apprentissage automatique, d'intelligence artificielle, la statistique descriptive et l'interrogation relationnelle. L'objectif a été de concevoir des requêtes "data science" en utilisant des environnements classiques comme Kaggle, Colab, Notebooks Azure mais surtout des technologies émergentes, les dits Machine Learning Studios, proposés par exemple par Azure ML gallery studio, Databricks MLFlow, Google IA Studio. Ce stage a contribué à l'élaboration de tutoriels[8] concernant le langage Cypher – en particulier dans le cadre des requêtes data-science.

[6] `http://www.info.univ-tours.fr/ICVL/`

[7] Campagne d'évaluation sur la fouille de textes en français `https://deft.limsi.fr/2020/`. Cette participation a donné lieu à la publication d'un article (Minard et al., 2020).

[8] `https://gevargas.github.io/doing-studio/Neo4J-datascience-tutorial.html`

En plus des deux stages, le travail de doctorat de Nicolas Hiot, en partenariat avec l'entreprise Ennov[9], a donné lieu à une publication[10] dans le *workshop* DOING@ADBIS-TPDL-EDA. Ce travail s'insère complètement dans la thématique DOING et l'entreprise Ennov participe à nos activités via le coencadrement de Nicolas.

# 4 Prototypes en marge des activités DOING

Dans le cadre d'une collaboration entre le LIG France, le Birla Institute of Technology, India et le National Institute of Genetic Engineering and Biotechnology, Iran, deux prototypes ont été développés au LIG. Les données utilisées sont celles exportées par l'Allen Institute sur la COVID_19 pendant la pandémie :

- S_COVID engine[11] for exploring COVID-19 documents (Allen Institute Data Collection on COVID-19) un système d'exploration des données guidée par la personne basé sur des data science *pipelines*. Il propose un environnement d'exploration de données rassemblant différentes techniques d'interrogation telles que *query by example* et de *morphing* de requêtes pour faciliter l'exploration des données.
- CMTA: A framework for Multilingual COVID-19 Tweet Analysis (Infodemic Twitter Dataset, COVID-19: Poynter Resources). CMTA utilise un modèle d'apprentissage approfondi pour la détection et la classification de la désinformation et des émotions dans les tweet multilingues. Le CMTS utilise le BERT multilingue pour extraire des caractéristiques des données textuelles multilingues, qui sont ensuite classées dans une classe spécifique d'émotion et de désinformation.

S_COVID a été présenté par un des auteurs lors d'un Webinar DOING. Deux publications sont en cours d'évaluation.

# 5 Les contacts

## 5.1 Contacts avec autres ateliers et actions MADICS

Suite aux suggestions que nous avons reçu de l'équipe MADICS, nous avons pris contact avec l'atelier MADICS-HN, tout d'abord pour leur proposer une intervention dans le cadre de notre journée d'étude en avril, puis suite à son annulation, pour s'entretenir avec nous lors d'une réunion par visio le 16 octobre 2020. L'atelier MADICS-HN était représenté par Fatiha Idmhand et Sabine Loudcher. Cet échange nous a permis de mieux comprendre les axes de recherche de leur atelier. Nous pensons que des échanges au cours de journées d'étude

[9] `https://fr.ennov.com/`

[10] Joshua Amavi, Mírian Halfeld Ferrari, Nicolas Hiot: Natural Language Querying System Through Entity Enrichment. ADBIS/TPDL/EDA Workshops 2020: 36-48

[11] `https://github.com/MehrdadFarokhnejad/S_COVID`

seraient intéressants, mais nous ne pensons pas qu'il soit pertinent de proposer une action ensemble.

Nos deux ateliers s'intéressent aux données textuelles, mais MADICS-HN questionne les données comme objet de leur recherche et sur 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. Leurs objectifs, bien que d'intérêt aussi pour nous, ne constituent pas les activités centrales de DOING.

L'atelier MADICS-HN nous a invité à un événement organisé le 25 novembre 2020 sur la notion d'hétérogénéité des données et nous prévoyons de participer.

## 5.2 Contacts avec des partenaires nationaux et internationaux

Comme déjà mentionné lors de nos commentaires évaluant les *webinars* DOING, nous avons pu, pendant cette année, identifier des **collègues** intéressés et intéressants qui peuvent amener des points de vues riches et divers à notre groupe thématique. Dans ce cadre, nous pouvons citer, dès maintenant, Salima Benbernou (LIPADE) avec qui un contact plus étroit commence à être établit, Vasiliki Foufi (Université de Genève), avec qui nous voulons proposer des échanges via des stages doctorants (hors crise sanitaire, hélas), André Freitas (University of Manchester), dont les thématiques de recherche concernent à la fois le TAL et les bases de connaissances, et Laure Berti-Équille (IRD) dont le travail semble pouvoir nous apporter beaucoup de savoir-faire dans le cadre des requêtes *data science*.

Nous pensons que, en plus de discussions abordant nos enjeux thématiques au sens plus large, DOING peut contribuer à la constitution de sous-groupes de travail sur des aspects plus précis. Le travail en sous-groupes serait source d'un dynamisme entre les membres qui pourraient ensuite venir partager leurs discussions avec notre assemblé plus large. Cette méthode de travail nous semble enrichissante et prometteuse. Nous aimerions la concrétiser d'avantage.

Nous avons aussi entamé une procédure d'échanges avec des **spécialistes dans le domaine de la santé**. Par des coïncidences personnelles, nous avons fait connaissance avec des médecins brésiliens de l'*Universidade Federal de Juiz de Fora* (UFJF) intéressés à participer à DOING. Ces premiers contacts nous ont permis déjà d'envisager des directions nouvelles qui devraient guider nos réflexions à l'avenir. Il s'agit ici de recenser les besoins des spécialités pour construire un cadre qui servira de guide à l'extraction d'information. La collaboration première avec ces médecins pourrait nous permettre d'accèder à un ensemble de questions intéressantes, utiles dans leurs activités pour ensuite nous aider dans la classification des questions. Nous imaginons qu'un tel cadre peut inspirer la sélection de l'information à extraire ainsi que les méthodes d'analyse de données à rendre prioritaire lors de la proposition des requêtes *data science*. Nous sommes actuellement en contact avec deux médecins de l'hôpital univer-

sitaire[12] de l'*Universidade Federal de Juiz de Fora*, plus précisément, les chefs des services d'ophtalmologie et d'oto-rhino-laryngologiste. Un travail en amont, rendant plus clair les besoins dans nos domaines différents, est à entamer. Pendant cette étape, selon nos contacts, il serait possible que d'autres spécialistes de l'hôpital universitaire s'intéressent à nos activités. Nous voyons ici non seulement une opportunité d'échange avec les spécialistes de la santé mais aussi une ouverture internationale très intéressante qui vient renforcer les liens que nous avons déjà avec le Brésil (voir nos collaborations avec l'UFPR et l'UFRN[13]).

Autre opportunité, cette fois dans le domaine de l'environnement, est la proposition de l'ARD (Ambition Recherche et Développement) de Centre-Val-de-Loire JUNON[14] visant la création d'un pôle de recherche numérique sur l'environnement continental. Ce projet est en cours d'évaluation et les membres LIFO de DOING y sont impliqués. Une ouverture vers les **échanges avec les collègues chercheurs dans le domaine de l'environnent** est donc en route.

# 6 Bilan scientifique: vers la proposition d'une action

Les différentes discussions scientifiques que nous avons pu avoir autour des thématiques DOING nous ont permis de repérer certains cadres prioritaires.

En premier lieu, nos premiers contacts avec des spécialistes, nous font croire au besoin de guider l'étape d'extraction d'information – et donc la construction d'une base de données graphe – par les objectifs des consultations souhaitées en aval. L'idée ici est d'extraire uniquement les types d'information qui contribueront directement ou indirectement aux analyses et questions envisagées dans l'étape d'interrogation intelligente de la base. Ce constant introduit une originalité à nos discussions, car il impose un travail préalable, avec les spécialistes, pour recenser et classer les types de questions et les analyses visées. La discussion sur les méthodes d'extraction d'information en prenant en compte ce nouveau point de vue devrait enrichir nos échanges.

Un autre constat vient du fait que les collègues ayant manifesté le plus d'intérêt par les discussions au sein de DOING émanent du domaine de bases de données et font donc leur recherches dans ce cadre (il s'agit ici de penser aux contraintes non seulement comme des inférences, à proposer des approches de mises à jour en gardant la sémantique du monde fermé...). Autrement dit, les questions liées au web sémantique ne sont pas le cœur de leur réflexions (même si parfois le RDF est utilisé comme illustration de leur propos). L'idée nous est donc parvenue d'approfondir nos discussions et échanges avec ceux qui travaillent sur les graphes de propriétés et les langages de requêtes type

[12] `http://www2.ebserh.gov.br/web/hu-ufjf`

[13] Collaboration avec Martin Musicante, membre DOING, ancien collaborateur du LIG et du LIFO.

[14] Cité dans le rapport d'activités disponible à `https://www.poledream.org/wp-content/uploads/2020/07/RA-DREAM2019_BD22.pdf`

Cypher. Nous sommes particulièrement intéressés par les connexions que nous pourrions avoir avec le projet Open Cypher et GQL. Nous pensons aussi qu'il serait intéressant de contacter des collègues des équipes GDD (LS2N, Nantes), Wimmics (Inria, Sophia Antipolis et Université Côte d'Azur) et BD (LIRIS, Lyon) dans le cadre du projet ANR DeKaloG. En effet, leur construction et leur exploitation des graphes de connaissances se rapprochent de nos réflexions sur la construction des bases de données graphes et leur interrogation avec des requêtes data science.

Ces deux directions 'prioritaires' ne restreignent pas nos discussions, mais, elles permettent au contraire d'établir les bases pour bâtir les interfaces nécessaires à la collaboration entre les spécialistes dans le domaine de connaissance (principalement les médecins), les chercheurs en traitement automatique des langues, en bases de données et en intelligence artificielle.

Nous croyons qu'il s'agit d'un premier pas vers une organisation plus ciblée de nos discussions, avec des passerelles vers des thèmes à la fois très proches et cependant distincts par leurs sensibilités et leurs préoccupations. Donner priorité à un cadre et définir ces passerelles seraient un premier pas vers la structuration d'une communication et d'une interaction plus productive avec d'autres actions ou ateliers.

# 7 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`

# 8 Bilan financier

Nous avons prévu le financement d'organisation de journées d'études et l'invitation de collègues qui pourraient présenter des conférences invitées.

Dans notre plan d'activités nous avions prévu l'organisation d'une première journée d'études à Paris (INALCO) en prenant en charge une pause-café et un buffet dinatoire. Nous avions prévu le financement partiel de la mission de Vasiliki Foufi (trajet Genève Paris, 1 nuit d'hôtel). Une deuxième journée d'étude dans la même configuration était prévue pour le deuxième semestre de l'année 2020.

A cause des mesures sanitaires, toutes nos réunions ont été réalisées de manière virtuelle et nous n'avons pas utilisé le budget accordé à DOING. Nous avons donc décidé de faire remonter ce budget à MADICS.

# 9 Publications associées à DOING


1. Anne-Lyse Minard, Andréane Roques, Nicolas Hiot, Mirian Halfeld Ferrari Alves, Agata Savary. *DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées.* 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), 2020, Nancy, France. pp.66-78.

2. Joshua Amavi, Mírian Halfeld Ferrari, Nicolas Hiot: *Natural Language Querying System Through Entity Enrichment.* ADBIS/TPDL/EDA Workshops 2020: 36-48

3. Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Javier-Alfonso Espinosa-Oviedo, *Enacting Data Science Pipelines for Exploring Graphs: From Libraries to Studios.* ADBIS/TPDL/EDA Workshops 2020: 271-280

4. Genoveva Vargas-Solar, *Are Data Science Pipelines Fuzzy Queries?* In Proceedings of the FASSAE Workshop, in conjunction with the International Conference on HPCS, 2020 (to appear)

5. Ladjel Bellatreche, Fadila Bentayeb, Mária Bieliková, Omar Boussaid, Barbara Catania, Paolo Ceravolo, Elena Demidova, Mírian Halfeld Ferrari, María Teresa Gómez López, Carmem S. Hara, Slavica Kordic, Ivan Lukovic, Andrea Mannocci, Paolo Manghi, Francesco Osborne, Christos Papatheodorou, Sonja Ristic, Dimitris Sacharidis, Oscar Romero, Angelo A. Salatino, Guilaine Talens, Maurice van Keulen, Thanasis Vergoulis, Maja Zumer: *Databases and Information Systems in the AI Era: Contributions from ADBIS, TPDL and EDA 2020 Workshops and Doctoral Consortium* ADBIS/TPDL/EDA Workshops 2020: 3-20

6. Jacques Chabin, Cédric Eichler, Mirian Halfeld-Ferrari, Nicolas Hiot, *Graph Rewriting Rules for RDF Database Evolution Management*, The 22nd International Conference on Information Integration and Web-based Applications & Services (iiWAS2020), to appear

7. Jacques Chabin, Cristina D. A. Ciferri, Mirian Halfeld-Ferrari, Carmem S. Hara, and Raqueline R. M. Penteado, *Role-Based Access Control on Graph Databases*, 47th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM) 2021, to appear.