WorkShop MoDaL (Multi-Scale Data Links)

Aurélien Cornet, Olivier Dameron, Alban Gaignard, Camille Maumet, Richard Redon, Anne Siegel

Atelier MoDaL (Multi-Scale Data Links)

Aurélien Cornet, Olivier Dameron, Alban Gaignard,
Camille Maumet, Richard Redon, Anne Siegel

28 janvier 2020

Table des matières

1 Introduction **2**

2 Exposés invités **2**

2.1 *Injecting Domain Knowledge in Electronic Medical Records to Improve Hospitalization Prediction*, Catherine Faron Zucker . . . 2

2.2 *INEX-MED: knowledge graphs to bridge imaging and omics data*, Alban Gaignard . . . 2

2.3 *RDF-datahub for precision medicine*, Olivier Dameron . . . 3

2.4 *Towards an Integrated Environment for Neuroscience Research*, Ivan Moszer . . . 3

2.5 *Bridging neuroimaging standards with JSON-LD*, Natacha Perez . . . 3

2.6 *Storing neuroimaging data for research: the experience of the Neurinfo platform*, Isabelle Corouge . . . 4

2.7 *The ins and outs of the Shanoir database*, Julien Louis . . . 4

3 Table ronde **4**

3.1 Gestion des données . . . 4

3.2 Ontologies . . . 5

4 Conclusion et perspectives **5**

1 Introduction

MoDal est un projet fédérateur inter-régions financé par Biogenouest, et porté par Christian Barillot, Anne Siegel, Olivier Dameron et Camille Maumet d'IRISA Rennes ainsi que Richard Redon et Alban Gaignard de l'institut du thorax à Nantes.

Les recherches explorant le domaine du vivant sont confrontées de manière croissante au besoin de devoir lier des données de phénotypage et de génotypage. Ces données de génomique et d'imagerie in-vitro, in-vivo sont aujourd'hui pour la plupart exploitées "en silos", c'est-à-dire sans moyen de pouvoir réaliser des analyses fines sur leurs complémentarités. Malgré les initiatives de mutualisation et de standardisation au travers de grandes infrastructures de recherche (FLI, FBI, IFB), y compris entre les instituts de recherche Inserm, CNRS, Inria et les Universités, il est aujourd'hui très difficile d'envisager une exploitation algorithmique et statistique conjointe de ces diverses sources de données.

En s'appuyant sur un cas d'usage réel et d'envergure nationale en santé (projet ICAN, IntraCranial ANeurysms), le projet fédérateur MoDaL (Multiscale Data Links) vise à décloisonner les ressources dédiées à l'imagerie et à la génétique. MoDaL s'articule autour de i) l'établissement d'un état des lieux des acteurs et des infrastructures disponibles à l'échelle de l'inter-région, ii) la proposition de démonstrateurs technologiques mettant en jeu des problématiques soulevées par ICAN et adressant la gestion, l'analyse et la réutilisation de données multi-infrastructures (imagerie in-vivo, in-vitro, génomique).

Ce document résulte du premier séminaire MoDal tenu le 11 juillet 2019. Il propose un compte rendu des présentations du séminaire, et une synthèse des enjeux et verrous communs résultant des échanges.

2 Exposés invités

2.1 *Injecting Domain Knowledge in Electronic Medical Records to Improve Hospitalization Prediction*, Catherine Faron Zucker

Catherine Faron est maître de conférence à l'Université de Nice, et travaille au sein de l'équipe Wimmics au laboratoire I3S de Sophia-Antipolis, sur la thématique "Linking Natural and Artificial Intelligence on the Web", faire le lien entre sémantiques formelles et sémantiques socials sur le web. L'enjeu est de prédire le risque d'hospitalisation par la détection de "facteurs de risque" dans les données patient, afin de rediriger ce patient chez des spécialistes et éviter l'admission ainsi que les coûts qui y sont liés. Ses recherches visent à utiliser les technologies du web sémantique en interaction avec le machine learning afin de faire de la médecine de précision. L'objectif est d'améliorer les prédictions en injectant des connaissances expertes (base de données publiques) et formelles (données cliniques). En principe, des "bag of words" sont récupérés à l'aide d'un algorithme à partir des données patients (base de données PRIMEGE). Ils sont ensuite traduits en concepts (ontologies) et croisés à des troubles ou maladies à l'aide de DBpedia et Wikidata.

2.2 *INEX-MED: knowledge graphs to bridge imaging and omics data*, Alban Gaignard

Alban Gaignard est ingénieur de recherche à l'institut du Thorax à Nantes sur le projet Inex-Med financé par l'institut français de bionformatique IFB et co-porté par Julie Thompson de l'Icube à Strasbourg. Le projet s'appuie sur les graphes de connaissance et l'apprentissage automatique pour lier et exploiter algorithmiquement des données hétérogènes (annotations cliniques, imagerie, génotypages, exomes). Le jeu de données est basé sur l'étude ICAN, une initiative nationale sur les anévrismes intracrâniens. Inex-Med représente toutes ces données dans un graphe typé et orienté, et permet par exemple de lier à un individu des informations sur l'angle entre deux bifurcations de

vaisseaux. L'objectif à terme est de faire des requêtes SPARQL sur le graphe pour extraire et apprendre des motifs intéressants.

2.3 *RDF-datahub for precision medicine*, Olivier Dameron

Olivier Dameron est maître de conférence à l'université de Rennes 1, et travaille au sein de l'équipe Dyliss sur le projet RDF-Datahub. Ce projet porte sur l'utilisation du web sémantique au service de la médecine de précision, qui consiste à sélectionner une solution optimale basée sur le profil génétique, cellulaire et moléculaire du patient. Cela demande de croiser des données hétérogènes, parfois massives, de différentes bases de données. Cela soulève également des problèmes comme le temps de réponse des requêtes. Une solution proposée pour éviter les "time out" (temps attribué à une tâche dépassé) consiste à découper les requêtes, envoyer les différents fragments à des endpoints (serveurs ou sources de données) ciblés susceptibles d'avoir l'information recherchée, puis faire l'union. En réduisant le nombre d'endpoints à passer en revue, le temps de recherche est diminué.

2.4 *Towards an Integrated Environment for Neuroscience Research*, Ivan Moszer

Ivan Moszer travaille en tant que coordinateur de la plateforme de bioinformatique ICONICS de l'institut du cerveau et de la moelle épinière ICM, situé à l'hôpital de la Pitié-Salpêtrière à Paris. Les thèmes de recherche qui y sont abordés comprennent la médecine de précision, l'aide à la décision, la modélisation de maladies, la translation de l'animal à l'homme. Afin de traiter, stocker et analyser les données nombreux types de données issus des recherches, différents outils sont nécessaires. Ivan en présente quelques uns :

- REDCap, (Research Electronic Data Capture), permet la création et la gestion de bases de données ainsi que d'études et d'enquêtes principalement dans le domaine de la recherche clinique.
- XNAT est une plateforme dédiée à la recherche en neuroimagerie.
- OMERO est un outil centré sur l'analyse et la gestion de données liées à la microscopie.
- tranSMART (Datawarehouse for translational research data), permet de stocker de larges quantités de données cliniques.

Enfin, Ivan travaille sur le projet "Data Lake", un espace unique de stockage pour toutes les données de l'ICM, brutes ou traitées, internes ou publiques. Le but est de croiser toutes ces données hétérogènes afin de faciliter la découverte de nouveaux biomarqueurs.

2.5 *Bridging neuroimaging standards with JSON-LD*, Natacha Perez

Natacha Perez travaille en tant que stagiaire au sein de l'équipe Empenn à l'IRISA de Rennes sur l'union de deux standards de données : BIDS et NIDM. Ces deux formats visent à proposer une structuration et un langage commun concernant les données en neuroimagerie. Il existe énormément de bases de données en neuroimagerie, et autant de structures et de nommages différents pour chacune. BIDS est une initiative proposant une structure de données pour la neuroimagerie, elle est déjà adoptée par une grande partie de la communauté et donne accès à plus de 20 000 sets d'imagerie. Le modèle de données NIDM propose un encodage des résultats d'analyses de neuroimagerie pour une représentation unifiée. Les deux formats ont un objectif similaire, mais BIDS ne se concentre que sur les données brutes, tandis que NIDM aspire à couvrir tous les types de données en neuroimagerie. De plus NIDM permet de faire le lien avec le web sémantique pour injecter ces données dans des graphes de connaissances.

2.6 *Storing neuroimaging data for research: the experience of the Neurinfo platform*, Isabelle Corouge

Isabelle Corouge travaille sur la plateforme de neuroimagerie NeurInfo en tant qu'ingénieur de recherche. Les objectifs de la plateforme Shanoir sont de promouvoir et aider la recherche dans le domaine des neurosciences, depuis des questions cliniques jusqu'à l'exploitation et le stockage des données. La majeure partie des données proviennent du CHU de Rennes. La plateforme développe la base de données Shanoir, une initiative nationale soutenue par l'infrastructure nationale de recherche FLI (France Life Imaging). Elle offre un espace de stockage pour archiver, indexer et partager des données issues d'étude cliniques, comme de l'imagerie cérébral. La base de données Shanoir permet de répartir des droits d'accès par études. Chaque étude peut contenir différentes images et techniques d'acquisition, ainsi qu'un large panel de métadonnées.

2.7 *The ins and outs of the Shanoir database*, Julien Louis

Julien Louis est ingénieur à l'IRISA Rennes et fait partie de l'équipe de développement de Shanoir. En 2016, la décision a été prise de refondre l'application car elle s'appuyait sur des librairies obsolètes. La nouvelle version, toujours en cours de développement, permet l'interopérabilité, supporte mieux la charge des données et est plus sécurisée.

3 Table ronde

Les interventions ont fait référence à plusieurs organismes, comme des travaux sur l'homme à des fins biomédicales, la prise en charge des modèles animaux (notamment dans Shanoir, présentation par Isabelle Corouge) ainsi que les modèles marins (Patrick Durand IFREMER). Plusieurs échelles d'études ont également été abordées comme les échelles cellulaire et moléculaire (avec une intervention de Gwénaël Rabut sur les données d'interactions protéines-protéines chez la levure). Parmi ces domaines de recherche très différents mais tous centrés sur 'les données', les enjeux qui se détachent tournent principalement autour de la mise en commun de données hétérogènes. Les principaux types de données qui ressortent sont l'imagerie médicale avec un focus sur la neuroimagerie, les données patients de phénotypage et observations cliniques, et les données génomiques.

En terme de solutions ou d'outils mis en place pour répondre à ces enjeux, on retrouve des solutions de stockage pour des données issues d'études cliniques (Shanoir, tranSMART), des outils d'analyses (OMERO en microscopie). On retrouve également les technologies du web sémantique, l'idée étant d'intégrer toutes les données hétérogènes au sein d'un graphe de connaissance (au format RDF) pour ensuite faire ressortir des biomarqueurs d'intérêt à l'aide de requêtes SPARQL. L'apprentissage machine a également été proposé sur ces graphes et détecter ces biomarqueurs. Des verrous existent, comme les limitations des ontologies existantes, dédiées pour décrire les entités du graphe de connaissance. Une ontologie commune permettrait un partage des données facilité. La scalabilité est aussi un problème, le temps de recherche augmente exponentiellement avec la complexité des requêtes et la taille de graphes, et peut renvoyer un time-out.

Les échanges avec les participants se sont concentrées sur deux grands points : la gestion de données et les ontologies

3.1 Gestion des données

Le stockage et le traitement des données est un point clé quelque soit le domaine abordé. Les solutions mises en place sont dans la grande majorité à l'échelle d'un institut, ou d'un groupe de collaboration travaillant sur un sujet d'étude précis (comme le projet de datalake de l'ICM). La création d'un espace de stockage est toujours motivé par un objectif, ce qui implique que la structure de la base de données et des outils sont développés précisément pour y répondre. Or les enjeux de recherche actuelle demandent de plus en plus un traitement croisé de différentes sources, de

différents types de données, de différents domaines, parfois massives et à des échelles différentes. Il faut rendre plus explicite les schémas et structures de bases de données et de connaissances.

Cela pose de nouvelles questions sur la structure de telles bases de données. Le critère de recherche principal qui fait office de point d'entrée diffère entre un clinicien cherchant un dossier patient, et un biologiste travaillant sur des protéines. L'aspect temporel est également différent, les données n'ayant pas le même cycle de vie entre les différents domaines de la biologie. Le problème intervient également dans le cas de la protection des données, ne concernant que les données humaines.

3.2 Ontologies

Il existe aujourd'hui peu d'ontologies standardisées pour faciliter les échanges et le croisement des données entre les domaines, le principal exemple étant BridgeDB. Des initiatives existent au sein d'un domaine, c'est le cas notamment en neuroimagerie (par exemple NIDM). Dans l'ensemble les données brutes sont correctement annotés, ce qui n'est pas le cas pour les données d'analyses traitées. Il existe beaucoup d'ontologies, la plupart développée en parallèle d'un outil pour répondre à un problème précis. Cela rend le traitement et le partage de données difficile, et ce même au sein d'un même institut. Développer des ontologies communes permettrait de gagner du temps, faciliterait le partage des données intra et inter-institut et limiterait les pertes de données. L'idée générale derrière est de favoriser l'interopérabilité.

4 Conclusion et perspectives

Cette journée de rencontre autour du projet MoDal a permis d'obtenir un premier état des lieux des verrous présents dans les différents domaines de la biologie présents. Certains sont spécifiques, comme la protection des données en santé, d'autres sont connus mais difficilement adressables, comme le passage à l'échelle technique pour gérer de grandes quantités de données. Néanmoins, les questions autour de l'interopérabilité des outils, du partage de données hétérogènes et des ontologies ont trouvé écho dans tous les domaines présentés. Cela ouvre des premières pistes à approfondir dans le cadre du projet MoDal. Les objectifs sont multiples :

- Poursuivre l'animation de la communauté en précisant ces verrous communs, notamment au cours d'interviews.
- Étendre cette communauté aux domaines peu représentés lors de cette journée, comme le végétal et le marin.
- A partir des problématiques collectivement identifiées, proposer et implémenter un ou plusieurs démonstrateurs technologiques.