INEX-MED – INtégration et EXploration de données bioMEDicales hétérogènes

Alban Gaignard[1], Julie Thompson[2], Kirsley Chennen[2,3], Maxime Folschette[4,5], Jocelyn Laporte[3], Olivier Poch[2], Richard Redon[1], Hala Skaf-Molli[5], et l'ensemble du consortium INEX-MED

[1]Institut du Thorax, Inserm UMR 1087, CNRS UMR 6291, Université de Nantes
[2]ICube, CNRS UMR 7357, Université de Strasbourg
[3]IGBMC, Institut de Génétique et de Biologie Moléculaire et Cellulaire, INSERM U1258, CNRS UMR 7104, Université de Strasbourg
[4]IFB, Institut Français de Bioinformatique, CNRS UMS 3601
[5]LS2N, Laboratoire des Sciences du Numérique de Nantes, CNRS UMR 6004, Université de Nantes

• ***Knowledge Graphs*** • ***Linked Data*** • ***FAIR Data*** • ***Machine Learning*** • **RDF / SPARQL**

1. Introduction

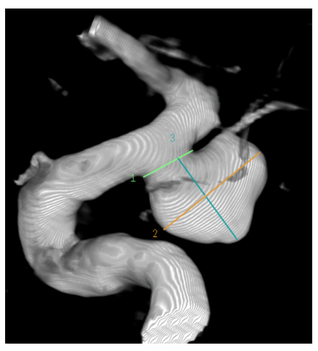
La médecine moderne nécessite le développement d'**approches interdisciplinaires**, pour exploiter des **données en grand nombre, multi-échelles, multi-sources**, issues des nouvelles modalités de phénotypage (par ex. imagerie) ou bien des nouvelles technologies "OMIC" (par ex. exomes, génomes). Mais l'intégration et l'analyse de ces nouvelles collections de données reste un défi majeur.

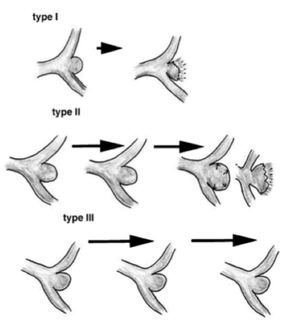
En s'appuyant sur les principes **FAIR data** [1] (*Findability, Accessibility, Interoperability, Reusability*), INEX-MED permettra d'identifier les verrous d'**interopérabilité** entre ces sources de données multi-échelles et des infrastructures de calcul multi-sites au travers de deux démonstrateurs. Ceux-ci visent à faciliter l'exploitation intégrée et la modélisation (sémantique/statistique) dans le contexte de deux cas d'études : les **anévrismes intracrâniens** et les **myopathies congénitales**.

Les compétences et expériences complémentaires des partenaires permettront de développer des solutions informatiques (bases de connaissances, interfaces de requêtes, analyses statistiques et apprentissage automatique) réutilisables pour de futurs projets intégratifs dans le domaine biomédical.

2. Anévrismes intracrâniens

Les anévrismes intracrâniens touchent environ 3,2 % de la population mondiale, et on estime que 50 % des cas de rupture d'AIC sont mortels [2].





Objectifs

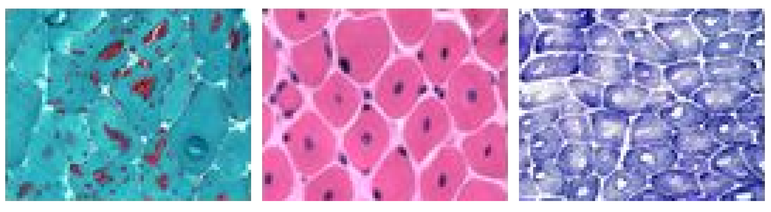
- Identifier des biomarqueurs quantitatifs d'**imagerie** spécifiques de patients porteurs de mutations **génétiques**
- Identifer des variants génétiques spécifiques de certains **phénotypes** observés en imagerie (par ex. mesures d'angles des bifurcations, tortuosité, localisation des anévrismes, etc.),
- Évaluer les facteurs d'exposition (habitude de vie) sur des données d'imagerie, etc.

Infrastructures

• **FLI** • **France Génomique** • **IFB**

3. Myopathies congénitales

Il existe différents sous-types de myopathies selon les anomalies du tissu musculaire. Le diagnostic moléculaire n'est concluant que dans moins de 50 % des cas (hétérogénéité clinique, génétique et histopathologique) [3].



à agrégation de protéines

centronucléaire

avec "cores"

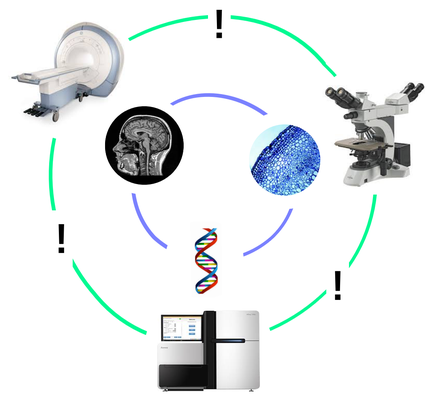
Objectifs

- Proposer une classification plus précise,
- Description des relations **génotypes-phénotypes**,
- Construire un modèle de diagnostic à partir des données phénotypiques à l'aide d'approches d'intelligence artificielle.

Infrastructures

• **CNRGH** • **BIOBANQUE** • **IFB**

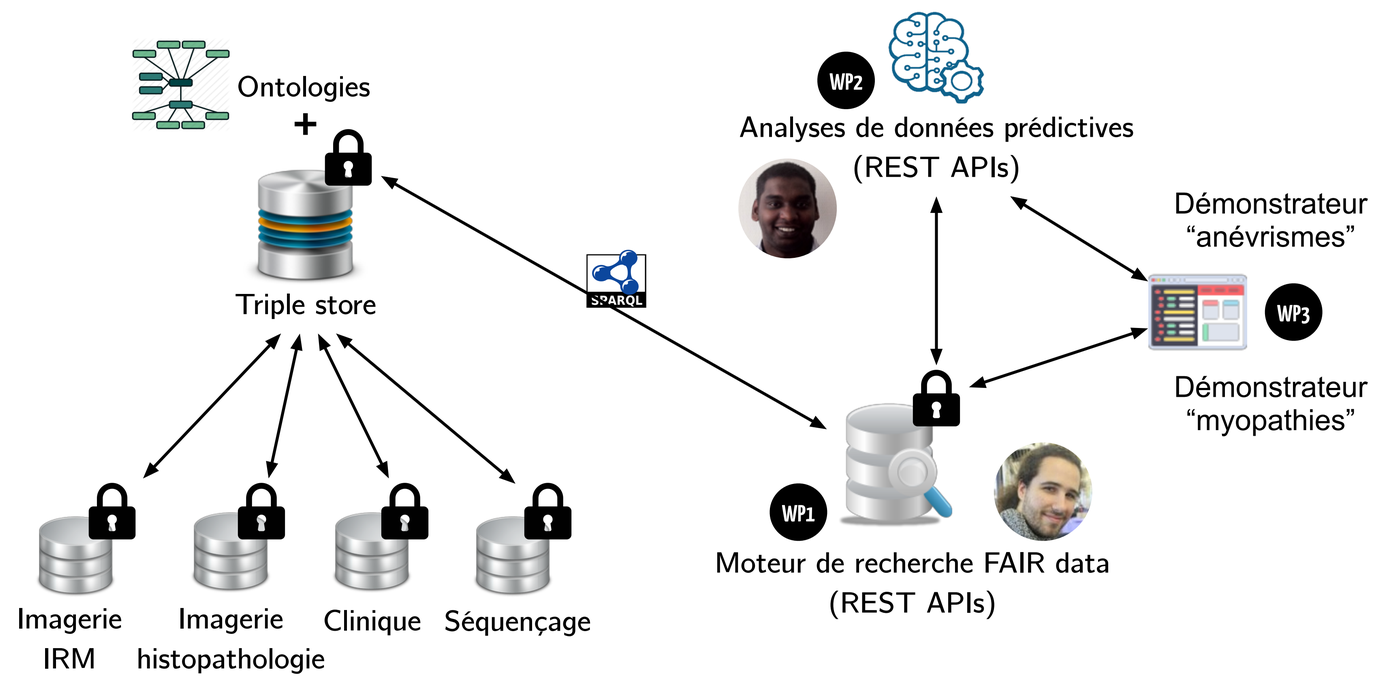
4. Défis



La recherche biomédicale fait aujourd'hui face à la disponibilité d'un très grand nombre de sources de données (données omics, réseaux, images, textes, etc.). Ces sources de données **hétérogènes et massives** proviennent d'observations sur des **échelles du vivant** extrêmement variées (gènes, cellules, organes, communautés, habitudes de vie, etc.). Mais elles restent cloisonnées au sein de "**silos**" spécifiques des communautés/pratiques scientifiques, ou répartis physiquement. Leur co-exploitation statistique et algorithmique nécessite l'utilisation d'un vocabulaire commun et reste aujourd'hui un défi.

• **Silos de données** • **Interopérabilité** • ***Linked Data***
• **Ontologies** • **variables (P) >> échantillons (N)**

5. Approche et résultats attendus



WP1 Unification des données (clinique, imagerie, séquençage, etc.) sous la forme d'un graphe de connaissances (*Linked Data*).
⇒ Infrastructure pour l'exploitation intégrée de données multidisciplinaires, à l'aide d'ontologies existantes ou nouvellement créées ; catalogue de requêtes (SPARQL), API REST, contrôle d'accès.

WP2 Exploitation de ce graphe de connaissances à l'aide de méthodes biostatistiques et d'apprentissage automatique.
⇒ Classifications et prédictions à l'aide d'algorithmes d'apprentissage ; détection de nouveaux marqueurs multi-source ; outils d'aide au diagnostic.

WP3 Développement de démonstrateurs web.
⇒ API REST avec contrôle d'accès et visualisation de résultats.

7. Références

[1] Mark D Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3 :160018, March 2016.

[2] Romain Bourcier et al. Understanding the Pathophysiology of Intracranial Aneurysm : The ICAN Project. *Neurosurgery*, 80(4) :621–626, 2017.

[3] J. Böhm et al. Integrated analysis of the large-scale sequencing project Myocapture to identify novel genes for myopathies. *Neuromuscular Disorders*, 27 :S195, 2017. 22nd International Congress of the World Muscle Society.