Proposition d'un sujet de thèse en cotutelle 2026/2027

A. **Directeur/rice de thèse à l'Université Libanaise**

1. Nom : MOHANNA
2. Prénom : Yasser
3. Titre (Prof, HDR, ….) : Prof.
4. Laboratoire : HKS LABORATOIRE FACULTE DES SCIENCES 1 UNIVERITE LIBANAISE HADATH
5. Adresse Web :
6. Etablissement :FACULTE DES SCIENCES 1 UNIVERITE LIBANAISE
7. Adresse Web :www.ul.edu.lb
8. Domaines d'expertise : traitement de signal et d'image, machine Learning, systèmes embarqués
   –
9. Publications importantes en relation avec le sujet proposé :
   –
   –
   –
10. Adresse Web de votre page personnelle :
11. Adresse e-mail : yamoha@ul.edu.lb

B. Directeur/rice de thèse à l'ISEN :

1. Nom : MOALIC
2. Prénom : Yann
3. Titre (Prof, HDR, …) : Enseignant-Chercheur, HDR
4. Laboratoire : LabISEN
5. Etablissement : ISEN YncréaOuest
6. Adresse Web :

   https://isen-ouest.fr/cv/yann-moalic/

7. Domaines d'expertise :
   - Computational biology
   - Network analysis
8. Publications importantes en relation avec le sujet proposé :
   - Biogeography revisited with network theory: retracing the history of hydrothermal vent communities. Moalic Y, Desbruyeres D, Duarte CM, Rozenfeld AF, Bachraty C, Arnaud-Haond S. 2012. Syst Biol 61:127-37.
   - Large expert-curated database for benchmarking document similarity detection in biomedical literature search. Brown P, Consortium R, Zhou Y. 2019.Database 2019.
   - Learning Ecological Networks from Next-Generation Sequencing Data, p 1-39. Vacher C, Tamaddoni-Nezhad A, Kamenova S, Peyrard N, Moalic Y, Sabbadin R, Schwaller L, Chiquet J, Smith MA, Vallance J, Fievet V, Jakuschkin B, Bohan DA. 2016. Chapter One -*In* Guy W, David AB (ed), Advances in Ecological Research, vol Volume 54. Academic Press.

9. Adresse Web de votre page personnelle :
10. Adresse e-mail : yann.moalic@isen-ouest.yncrea.fr

C. Description du sujet de thèse proposé : (3 à 5 pages)

1. Discipline :

   Sciences et technologies de l'information / Sciences marines et environnementales (interdisciplinaire : intelligence artificielle, ingénierie des connaissances, interopérabilité sémantique et science des données).

2. **Titre et résumé :**

Architecture agentique locale pour le traitement et la valorisation de données marines hétérogènes

Résumé : L'océan est observé par une mosaïque d'infrastructures produisant des données biologiques, physiques et biogéochimiques issues de plusieurs plateformes d'observation et de programmes scientifiques. Ces données bénéficient de différents niveaux de standardisations et d'interopérabilité sémantique mais restent difficiles à croiser en pratique : le chercheur doit connaître les schémas, les résolutions spatio-temporelles, les services et les vocabulaires de chaque infrastructure, ce qui constitue un frein majeur à la science interdisciplinaire et aux principes FAIR.

Cette thèse propose de concevoir une architecture d'agents IA associée à des graphes de connaissances RDF, capable (i) d'assister le chercheur dans le mapping de ses données brutes vers des schémas standardisés, mais également dans la vérification, l'harmonisation et l'enrichissement sémantique des données déjà structurées et

diffusées, en s'appuyant sur I-ADOPT (Magagna et al. 2022) et aux vocabulaires contrôlés ; (ii) d'exploiter ces correspondances sémantiques pour traduire des questions scientifiques formulées en langage naturel en requêtes croisées sur des sources de de données hétérogènes à forte composante spatio-temporelle, locales ou en ligne, et (iii) de générer automatiquement des visualisations certifiées (cartes, données discrètes, séries-temporelles et analyses comparatives) accompagnées d'un cahier de laboratoire numérique documentant la provenance et la conformité de chaque traitement. Le système sera évalué sur un cas d'usage pilote et par rapport à un benchmark reproductible. L'utilisation d'un ou plusieurs modèles agentiques locaux, que ce soit sur la machine du chercheur ou sur un serveur de son institution garantit la maîtrise complète de l'utilisation de ressources (notamment énergétiques), de ses données et de ses processus, puisque rien ne quitte l'infrastructure, garantissant ainsi confidentialité, conformité réglementaire et indépendance vis-à-vis des services cloud tiers.

3. Sujet :

La communauté marine s'est dotée, en vingt ans, d'un vaste écosystème d'infrastructures permettant de collecter, structurer et diffuser des données biologiques, physiques et biogéochimiques. Cette évolution s'est accompagnée de l'adoption de standards de description et d'échange des données (Darwin Core, conventions CF), de vocabulaires contrôlés pour les paramètres scientifiques (NERC Vocabulary Server – NVS) et de référentiels spécialisés, notamment taxonomiques (WoRMS). Malgré ces avancées, les données demeurent hétérogènes en raison de la diversité de leurs formats, modèles de représentation, résolutions spatio-temporelles et vocabulaires scientifiques. Leur harmonisation et leur exploitation conjointe nécessitent encore de nombreuses interventions manuelles, limitant leur réutilisation conformément aux principes FAIR (Wilkinson et al., 2016) et aux bonnes pratiques scientifiques promues par l'Ocean Best Practices System (Pearlman et al., 2019).

Parallèlement, l'émergence des grands modèles de langage (LLM) outillés et des architectures d'agents IA ouvre de nouvelles perspectives pour automatiser la découverte, l'interprétation et le croisement de données scientifiques. L'association de ces approches avec des graphes de connaissances RDF et des modèles de représentation sémantique, notamment I-ADOPT (Magagna et al., 2022), permet d'envisager des systèmes capables d'assister les chercheurs dans l'harmonisation des données et leur exploitation. Ces mêmes graphes peuvent également être étendus à la représentation des opérations scientifiques et de leur provenance, afin de

relier les données d'origine, les traitements effectués et les résultats produits. Toutefois, la fiabilité des correspondances proposées, la cohérence des traitements et la reproductibilité des résultats constituent encore des défis scientifiques importants.

Le gap : il reste à concevoir et à évaluer une architecture intégrée combinant (a) l'assistance bidirectionnelle à la structuration, à l'annotation et à l'harmonisation sémantique des données, depuis les observations brutes jusqu'aux données déjà diffusées, (b) l'interrogation et le croisement intelligents de sources hétérogènes, en tenant compte de leurs caractéristiques sémantiques et spatio-temporelles, et (c) la génération automatisée de visualisations scientifiques vérifiables, accompagnées d'un cahier de laboratoire numérique fondé sur la traçabilité RDF des données, des transformations et des contrôles de conformité. L'enjeu est de déterminer comment les graphes de connaissances peuvent renforcer la fiabilité des agents IA tout au long de cette chaîne de traitement scientifique.

Une attention particulière sera également portée aux possibilités d'exécution locale des agents IA, sur la machine du chercheur ou sur un serveur institutionnel, notamment à partir de modèles ouverts. Cette approche pourrait permettre de mieux maîtriser la confidentialité des données, les processus de traitement et l'utilisation des ressources informatiques, y compris leur consommation énergétique, tout en réduisant la dépendance aux services cloud tiers.

Description du problème

Trois verrous concrets freinent aujourd'hui la valorisation des données marines :

1. **L'harmonisation sémantique des données reste complexe et largement manuelle.** Le passage de données scientifiques brutes (tableurs de campagnes, fichiers d'instruments, observations environnementales) vers des modèles et schémas standardisés nécessite d'identifier les variables, leurs unités, leurs méthodes de mesure et leurs correspondances avec des référentiels scientifiques. Inversement, les données déjà structurées et diffusées peuvent présenter des incohérences, des ambiguïtés ou des annotations incomplètes. L'exploitation de vocabulaires contrôlés et de modèles sémantiques comme I-ADOPT (Magagna et al., 2022), associés à des graphes de connaissances RDF, offre des perspectives pour automatiser ces opérations tout en conservant la possibilité d'une validation scientifique.

2. **Le croisement de données hétérogènes demeure difficile à automatiser.** Répondre à une question scientifique impliquant plusieurs sources nécessite d'identifier les jeux de données pertinents, d'interpréter leurs variables, de résoudre les différences de représentations de données spatio-temporelles (descriptions en langue naturelle, formats, résolutions (Janowicz et al. 2012)), puis de définir des opérations adaptées de rapprochement, d'agrégation ou d'interpolation. Ces étapes mobilisent des compétences multiples et restent souvent réalisées à travers des traitements spécifiques, difficilement réutilisables. Le défi consiste à exploiter les correspondances sémantiques représentées dans les graphes de connaissances pour permettre aux agents IA de construire et d'exécuter des requêtes cohérentes, vérifiables et reproductibles.

3. **La traçabilité et la vérifiabilité des traitements scientifiques restent insuffisantes.** Les analyses et visualisations produites à partir de sources hétérogènes ne documentent pas systématiquement les versions des données, les transformations appliquées, les paramètres utilisés et les incertitudes associées. Cette difficulté est renforcée par l'utilisation d'agents IA, dont les décisions et les résultats doivent rester contrôlables et cohérents. **L'intégration de mécanismes de provenance RDF**, accompagnées d'un cahier de laboratoire numérique retraçant les opérations réalisées et les contrôles de conformité associés.

Objectifs opérationnels (3 piliers) :

- **P1 — Assistance à l'annotation et à l'harmonisation sémantique des données :** un agent interactif assiste le chercheur dans le mapping de ses données brutes vers des schémas standardisés, mais également dans la vérification et l'enrichissement sémantique des données déjà structurées et diffusées. Il s'appuie sur **I-ADOPT, les vocabulaires contrôlés et les graphes de connaissances RDF** pour identifier les variables, proposer des correspondances sémantiques, détecter les incohérences et faciliter leur validation par les utilisateurs.

- **P2 — Interrogation et croisement intelligent de données hétérogènes :** un agent exploite les correspondances sémantiques du premier pilier pour traduire une question scientifique formulée en langage naturel en un plan d'exécution : découverte des jeux de données, interrogation de sources locales ou distribuées, rapprochement des variables et croisement des observations selon leurs caractéristiques spatio-temporelles. Les traitements s'appuient sur des outils scientifiques adaptés, avec une attention particulière portée à la cohérence des résultats et aux incertitudes associées.

- **P3 — DataVis & cahier de labo automatique** : génération automatisée de visualisations scientifiques (cartes de distribution, séries temporelles, analyses comparatives) et export automatique d'un notebook/rapport attestant des contrôles de conformité des traitements : provenance de chaque donnée, paramètres de chaque étape, versions des agents et des modèles, journal d'exécution. Ces informations seront structurées et mises en relation afin d'assurer la cohérence et la continuité entre les données, les traitements et les résultats produits, pour constituer un « cahier de labo numérique » garantissant la traçabilité et facilitant la reproductibilité des résultats, conformément aux principes FAIR et aux Ocean Best Practices.

Cas d'usage pilote : le système sera expérimenté sur des jeux de données marines hétérogènes associant des observations biologiques, physiques et biogéochimiques. Les données et les questions scientifiques seront définies avec les partenaires afin d'évaluer les trois piliers sur une chaîne complète, depuis l'harmonisation sémantique jusqu'à la production de résultats vérifiables. Les performances seront comparées à celles d'approches conventionnelles à l'aide d'un benchmark reproductible.

a. Approche méthodologique :

- **État de l'art & spécification** (M1–M6) : revue systématique (agents LLM pour données environnementales, interopérabilité sémantique, I-ADOPT, graphes de connaissances RDF et provenance des traitements) ; formalisation d'un « graphe de tâches » reliant l'annotation sémantique, le croisement des données et leur visualisation ; définition du benchmark d'évaluation (ensemble de jeux de données hétérogènes et de questions scientifiques à réponses vérifiables, métriques : exactitude des correspondances sémantiques, justesse du matching

spatio-temporel, taux d'hallucination, complétude des annotations et reproductibilité des traitements).

- **Architecture agentique** (M6–M18) : LLM généraliste outillé via un registre d'outils typés (recherche de jeux de données, interrogation de graphes RDF, accès aux données CSV/NetCDF, services web, vocabulaires contrôlés, contrôle qualité et visualisation). Développement d'un graphe de connaissances fondé notamment sur I-ADOPT et les référentiels scientifiques existants. Orchestration agentique (agent planificateur et fonctions spécialisées selon les besoins), mécanismes de robustesse : contraintes de schéma, validation systématique des sorties, signalement explicite des incertitudes plutôt qu'approximation, journalisation des actions. Exploration de l'option « agent local/souverain » : modèles ouverts auto-hébergés, avec comparaison des performances, des ressources mobilisées et des contraintes d'exécution locale ou via API.

-

- **Assistant de conversion** (M12–M24) : développement interactif avec des chercheurs partenaires d'un assistant permettant le mapping des données brutes vers des schémas standardisés, mais également la vérification et l'enrichissement des données déjà structurées et diffusées. Exploitation des correspondances sémantiques pour faciliter l'interrogation et le croisement des observations. Évaluation utilisateur sur des jeux de données réels (précision des correspondances, temps de traitement, taux d'erreur et effort de validation par rapport aux méthodes manuelles).

- **Visualisation & certification** (M18–M30) : développement de pipelines de visualisation scientifique (cartes, séries temporelles, analyses comparatives) ; génération de notebooks reproductibles (Quarto/Jupyter) intégrant la provenance des données, les paramètres des traitements, les versions des outils et le journal d'exécution ; formalisation d'un modèle de métadonnées de traçabilité, aligné sur PROV-W3C, permettant de relier les données, les transformations et les résultats produits. Mise en place de contrôles explicites de conformité et de reproductibilité pour accompagner les visualisations d'un « cahier de labo numérique ».

- **Évaluation consolidée & diffusion** (M30–M36) : benchmark complet comparant les approches conventionnelles, les agents IA sans représentation sémantique explicite et l'architecture proposée ; tests avec des utilisateurs externes, publication des prototypes open source, valorisation scientifique et rédaction du manuscrit.

Positionnement éthique et reproductibilité : le code développé sera diffusé en open source (licence ouverte, dépôt public) et les jeux d'évaluation partageables seront publiés. L'approche agentique et les procédures de validation seront documentées selon les principes FAIR et les recommandations de l'Ocean Best Practices System (OBPS). L'exécution locale des agents restera une option technique, dont les bénéfices en matière de confidentialité, de maîtrise des ressources et d'indépendance technologique pourront être évalués.

b. Résultats attendus :

- **Un cadre formel d'interopérabilité sémantique** fondé sur I-ADOPT et les graphes de connaissances RDF, permettant l'alignement des variables scientifiques et le croisement de données marines hétérogènes, avec prise en compte de leurs caractéristiques et incertitudes spatio-temporelles (contribution scientifique principale, publiable).

- **Une architecture agentique open source** avec ses outils (SDK d'agents pour l'annotation sémantique, l'interrogation de sources hétérogènes, le contrôle qualité et la visualisation), accompagnée d'un **benchmark d'évaluation public** — contribution à l'ingénierie des connaissances et à l'exploitation intelligente des données environnementales.

- **Un assistant d'annotation et d'harmonisation sémantique opérationnel**, testé sur des jeux de données réels de laboratoires partenaires, permettant de structurer les données brutes, de vérifier et d'enrichir les données déjà diffusées, et de réduire les efforts nécessaires à leur mise en conformité avec les principes FAIR.

- **Des livrables de démonstration** : visualisations scientifiques vérifiables (cartes, séries temporelles, analyses comparatives), notebooks reproductibles, cahier de laboratoire numérique et rapports de conformité documentant la provenance et les transformations réalisées.

- **Publications visées (2–3)** : un article sur l'interopérabilité et l'alignement sémantique des données marines ; un article sur l'architecture agentique et son benchmark d'évaluation ; un article sur la génération de visualisations scientifiques, la traçabilité et la reproductibilité des traitements. Un *data paper* sur les jeux de données du cas d'usage pilote pourra également être envisagé selon les résultats obtenus.

c. Calendrier :

Période	Jalon
M1-M6	État de l'art, formalisation du problème, sélection des jeux de données et spécification du benchmark d'évaluation.
M6-M12	Premier prototype P1 : annotation sémantique, mapping des variables et construction du graphe de connaissances RDF/I-ADOPT.
M12-M18	Architecture agentique intégrée ; consolidation de P1 et début de P2 (requêtage intelligent et croisement de données hétérogènes).
M18-M24	P1 et P2 opérationnels, évaluations utilisateurs ; début de P3 (visualisations scientifiques et traçabilité des traitements).
M24-M30	P3 opérationnel : notebooks reproductibles, provenance PROV-O, contrôles de conformité ; benchmark consolidé et évaluation optionnelle de l'exécution locale des agents.
M30-M36	Évaluations externes, diffusion open source, publications scientifiques et rédaction du manuscrit.

d. Mots clés :

Agents IA ; LLM outillés ; données marines hétérogènes ; interopérabilité sémantique ; I-ADOPT ; graphes de connaissances RDF ; vocabulaires contrôlés ; harmonisation des données ; croisement spatio-temporel ; visualisation scientifique ; traçabilité et provenance ; cahier de laboratoire numérique ; FAIR ; reproductibilité ; Ocean Best Practices ; Open Science.

e. Possibilité de financement (Justificatif éventuel) :

f. Profil de candidats recherchés :
   - Master 2 avec de fortes compétences en informatique (génie logiciel, IA agentique, Optimisations)
   - Compétences appréciées fouille de données, modélisation de l'information, IA agentique, web sémantique, Systèmes d'Information Géographiques
   - Appétence pour les sujets en lien avec l'environnement marin

Bibliographie :

Janowicz, K., Scheider, S., Pehle, T., and Hart, G.: Geospatial semantics and linked spatiotemporal data – Past, present, and future, Semantic Web, 3, 321–332, https://doi.org/10.3233/SW-2012-0077, 2012.

Magagna, B., Moncoiffé, G., Devaraju, A., Stoica, M., Schindler, S., Pamment, A., and RDA I-ADOPT WG: InteroperAble Descriptions of Observable Property Terminologies (I-ADOPT) WG Outputs and Recommendations, version 1.1.0, Research Data Alliance, https://doi.org/10.15497/RDA00071, 2022.

Pearlman, J., Bushnell, M., Coppola, L., Karstensen, J., Buttigieg, P. L., Pearlman, F., Simpson, P., Barbier, M., Muller-Karger, F. E., Munoz-Mas, C., Pissierssens, P., Chandler, C., Hermes, J., Heslop, E., Jenkyns, R., Achterberg, E. P., Bensi, M., Bittig, H. C., Blandin, J., Bosch, J., Bourles, B., Bozzano, R., Buck, J. J. H., Burger, E. F., Cano, D., Cardin, V., Llorens, M. C., Cianca, A., Chen, H., Cusack, C., Delory, E., Garello, R., Giovanetti, G., Harscoat, V., Hartman, S., Heitsenrether, R., Jirka, S., Lara-Lopez, A., Lantéri, N., Leadbetter, A., Manzella, G., Maso, J., McCurdy, A., Moussat, E., Ntoumas, M., Pensieri, S., Petihakis, G., Pinardi, N., Pouliquen, S., Przeslawski, R., Roden, N. P., Silke, J., Tamburri, M. N., Tang, H., Tanhua, T., Telszewski, M., Testor, P., Thomas, J., Waldmann, C., and Whoriskey, F.: Evolving and Sustaining Ocean

Best Practices and Standards for the Next Decade, Front. Mar. Sci., 6, 277, https://doi.org/10.3389/fmars.2019.00277, 2019.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data, 3, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.