Management and integration of agronomical, phenotypical and environmental data

This mini-symposium is funded by the GDR CNRS MaDICS, the Phénome Infrastructure, and the INRAE BAP department.

Introduction - Fred de Lamotte pour le collectif d'organisation de GIDAPE

- Data, ontologies and resources (chair : Nicolas Parisey, Fabrice Legeai)

15h00 *Central role of metadata language to create a federated data endpoint with a high FAIRness degree.* **Yvan Le Bras**

15h25 Use of deep learning methods applied to the study of ticks and tick-borne diseases. J**ocelyn de Goër de Herve**

- New technologies for data collecting and harvesting (chair : Frédéric de Lamotte, Yvan Le Bras)

15h50 « Sicpa IoT » : un déploiement de capteurs environnementaux, à large échelle, en unités expérimentales animales à INRAE. **François Laperruque** (CATI SICPA INRAE, Toulouse)

16h15 PAUSE

- Data integration and Information System (chair : Cyril Pommier, Olivier Dameron)

16h45 PHIS and OpenSILEX information systems for FAIR data. **Gabriel Besombes, Valentin Rigolle**

17h10 *Rule Mining in Knowledge Graphs for Data integration and Beyond .* **Fatiha Saïs**

17h35-18h Discussion

Session 1 : Data, ontologies and resources (chair : Nicolas Parisey, Fabrice Legeai)

- *Central role of metadata language to create a federated data endpoint with a high FAIRness degree.* **Yvan Le Bras** (Infrastructure PNDB, MNHN, Concarneau).

The “PNDB”, the French Biodiversity Data hub, is an e-infrastructure created in 2018 and entering an operational implementation phasis for these tools and services. Through the 3 components data/metadata, animation and data analysis, I propose to illustrate these tools and services by focusing on the approach proposed to go from data to the use of terminological resources via the intensive use of a metadata standard with multiple capabilities: the Ecological Metadata Language, or EML!

- Use of deep learning methods applied to the stud of ticks and tick-borne diseases. Hossain[a,b,c], Yann Frendo[a,b,c], Valérie Poux[b,c], Isabelle Lebert[b,c], Delphine Martineau[d], Olivier Lesens[d,e], Engelbert Mephu Nguifo[a], Gwenaël Vourc’h[b, c], J**ocelyn de Goër de Herve**[b, c]

[a]Université Clermont Auvergne, CNRS, ENSMSE, LIMOS, F-63000 Clermont-Ferrand, France, [b]Université Clermont Auvergne, INRAE, VetAgro Sup, UMR EPIA, F-63122 Saint-Genès-Champanelle, France, [c]Université de Lyon, INRAE, VetAgro Sup, UMR EPIA, F-69280 Marcy l’Etoile, France, [d]Infectious and Tropical Diseases Department, CHU Clermont-Ferrand, F-63000 Clermont-Ferrand, France, [e]CHU Clermont-Ferrand, Inserm, Neuro-Dol, CNRS 6023 Laboratoire Microorganismes Génome Environnement (LMGE), Université Clermont Auvergne, F-63000 Clermont-Ferrand, France

Use of deep learning methods applied to the study of ticks and tick-borne diseases. Sk Imran
Ticks are important vectors of pathogens for human and animal populations in Europe. In a context of environmental changes, they are an Important source of concern for the population and a source of health risks that are frequently relayed by the public media. Therefore, surveillance and prevention are public health issues, which require the development of scientific research. These research projects are carried out in partnership with public research units, private partners or local collectivities. They involve observation systems (tick collection in the environment, observation and declaration of bites by citizens) and analysis of the data collected via statistical modelling, genomics or artificial intelligence methods.

In this context, we present two research projects illustrating the use of Artificial Intelligence (AI), in particular Deep Learning methods, that are applied to the problematics of ticks and Lyme disease:

The DAPPEM project (Development of an application for the Identification of Erythema Migrans) aimed to develop a smartphone application to detect the early signs of Lyme disease, called erythema migrans, from a photo and a context questionnaire. The detection is performed using a Convolutional Neural Network (CNN) that automatically analyzes photos taken by users and returns a risk score calculated together with the context questionnaire. In the case of a suspected erythema migrans, a message will guide the user to an appropriate care pathway. The application is currently in the process of certification as a medical device. It will be available in 2023, to healthcare professionals and to the general public. The data that would be collected anonymously will be used to improve the detection model and to contribute to statistical studies for estimating incidence of Lyme disease.

The DCLIC project (Deep Convolutional Learning Ixodidae Characterization) aims to develop a tool that can detect in real time ticks in a video and classify them within four taxonomic genus that are present in metropolitan France (*Ixodes, Hyalomma, Rhipicephalus, Dermacentor*) The works carried out during this project have demonstrated that it was possible to automate the identification of tick genus from photos or videos. The

latest version of the neural network is able to detect a tick from an image with an accuracy of 0.968 and a specificity of 0.963, when ticks are larger than 5% of the total surface of the photo. The methodological developments produced during the DCLIC project are intended to be shared with the scientific community and could be integrated into applications for research teams collecting ticks or into mobile applications for participatory action research.

Session 2 : New technologies for data collecting and harvesting (chair : Frédéric de Lamotte, Yvan Le Bras)

- « Sicpa IoT » : un déploiement de capteurs environnementaux, à large échelle, en unités expérimentales animales à INRAE. **François Laperruque** (CATI SICPA INRAE, Toulouse)

Le Cati SICPA (Systèmes d'Informations et Calcul pour le Phénotypage Animal) met en place des solutions de collecte, de transfert et de stockage des données, issues des unités expérimentales animales des départements Génétique Animale (GA) et Physiologie Animale et Systèmes d'Élevages. Sa composante calcul, contribue au déploiement de solutions pour l'analyse de données (infrastructure, logiciels, IA).

Dans le cadre du projet "Sicpa IoT", nous répondons aux demandes réglementaires et scientifiques d'enregistrements longitudinaux de données environnementales, en élevages. Les informations collectées sont importantes pour mieux comprendre l'expression des phénotypes des animaux (consommation d'eau, résistance aux maladies, comportement, ...). Le choix d'une infrastructure Lora, gérée en propre, a été réalisé. Celle-ci s'appuie sur un écosystème de standards ouverts et de logiciels libres, bénéficiant d'une large communauté d'utilisateurs.

Nous utilisons, par exemple, le serveur LoRaWAN Chirpstack, un broker MQTT pour transporter les données, ainsi que la base de données Time Series, InfluxDB pour leur stockage. L'ensemble de l'infrastructure s'appuie sur la solution OpenStack proposée par la DSI INRAE.

L'utilisation de standards ouverts va nous permettre de développer des interfaces simplifiées selon les types d'utilisateurs et de connecter des capteurs communicants à des formats divers.

Session 3 : Data integration and Information System (chair : Cyril Pommier, Olivier Dameron)

- PHIS and OpenSILEX information systems for FAIR data. Isabelle Alic, Anne Tireau, Arnaud Charleroy, **Gabriel Besombes, Valentin Rigolle** and Pascal Neveu (INRAE, Montpellier)

Research data are more and more complex and must evolve continuously to meet the scientific challenges of today and tomorrow. The information systems of research communities must therefore be able to handle this complexity and also meet the requirements of an open and reproducible science. In response to these challenges, we have developed OpenSILEX, a software suite that allows building information systems and implementing FAIR principles. These systems are ontology-driven.
OpenSILEX is derived under several implementations corresponding to specific scientific communities:

PHIS (high throughput plant phenotyping),

Sixtine (experimental plant units at INRAE),
Silex-VitiOeno (wine and viticulture),
Silex-LBE (depollution).

The PHIS implementation is the information system of the Phenome project. It proposes, beyond data management, a structured vocabulary for a better data sharing within the community.

- *Rule Mining in Knowledge Graphs for Data integration and Beyond .* **Fatiha Saïs** (LRI, Univ Paris-Saclay, Paris)

Abstract: Today, we are experiencing an unprecedented production of resources, published as Linked Open Data (LOD, for short). This is leading to the creation of knowledge graphs (KGs), such as DBpedia, YAGO and Wikidata on the academic side, and the Google Knowledge Graph or Microsoft's Satori graph on the commercial side. These KGs contain millions of entities (such as people, proteins, or books), and millions of facts about them. This knowledge is typically expressed in RDF (Resource Description Framework), i.e., as triples of the form ⟨Biden, presidentOf, USA⟩. Some KGs provide an ontology expressed in OWL2 (Web Ontology Language), which describes the vocabulary (the classes and properties) for the RDF facts. However, these knowledge graphs are far from being correct and complete. That is, the existing KGs need to be refined in order to be enriched while preventing introducing errors. In this talk, I will present some contributions for KG refinement task that we tackle through the mining of logical rules. These rules may be used for integrating heterogeneous information or to explain « causality » relationships between some phenomena. I will finish by highlighting some current challenges.