



École d'été interdisciplinaire en numérique de la santé du 3 au 7 juin 2024

Intégration et interrogation avancées de données et de connaissances grâce au Web sémantique

Plan d'activité pédagogique

Type de l'activité : ☒ Atelier ☐ Présentation

Objectif : Le présent document décrit le contenu scientifique et l'organisation de l'activité pédagogique « Intégration et interrogation avancées de données et de connaissances grâce au Web sémantique » présentée lors de l'école d'été.

1 Introduction

L'activité pédagogique « Intégration et interrogation avancées de données et de connaissances grâce au Web sémantique » se déroulera dans le cadre de l'école d'été interdisciplinaire en numérique de la santé (EINS) et s'inscrit dans le thème « Modélisation des connaissances et de données en santé ».

1.1 Renseignements sur le présentateur

Prénom, nom : Olivier Dameron

Affiliation principale :

Professeur à l'Université de Rennes (France), Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA).

Courriel : olivier.dameron@univ-rennes.fr

Site Web : <https://www-dyliss.irisa.fr/olivier-dameron/>

Biographie

Olivier Dameron développe des méthodes basées sur les ontologies pour analyser des données biomédicales. Cela fait intervenir des compétences en représentation des connaissances et en bio-informatique.

Son approche consiste à exploiter des connaissances symboliques du domaine d'étude afin d'améliorer l'analyse de données qui sont en grandes quantités, complexes, fortement interdépendantes et incomplètes. Il utilise les technologies du Web sémantique pour intégrer ces données, qui sont souvent distribuées, et pour combiner différents types de raisonnement : déduction, classification, comparaison...

L'application principale concerne la caractérisation fonctionnelle et la comparaison de voies métaboliques et de voies de signalisation.

Il est responsable de l'équipe DYLISS à l'IRISA. Olivier Dameron a été responsable du master 1 recherche « Méthodes et Traitements de l'Information Biomédicale et Hospitalière » de 2007 à 2012 et coresponsable du master de bio-informatique de l'Université de Rennes1 de 2012 à 2022.

2 Description

Cette section présente le contenu de l'activité et les principaux objectifs.

2.1 Contenu

Nous sommes dans une ère de production à grande échelle de données en sciences de la vie qui sont offertes sous forme électronique. Cependant, on s'aperçoit que le défi d'intégration des données et de leur interopérabilité devient de plus en plus complexe. Il n'est plus possible de se reposer sur des solutions *ad hoc*, et le besoin de méthodes et d'outils adaptés devient donc crucial. En effet, les données des sciences de la vie sont intrinsèquement compliquées — à cause du grand nombre d'éléments différents qui entrent en jeu —, et complexes — à cause de la forte interdépendance de ces éléments. La manière de traiter ces données est ainsi devenue un domaine d'étude à part entière.

Cet atelier s'appuiera sur les notions de base de développement des ontologies, en mettant l'accent sur l'ingénierie des données et des connaissances. Les grands principes de modélisation garantissant la bonne structuration des données en RDF (*Resource Description Framework*) et leur interopérabilité, en lien avec des bases de connaissances sous forme d'ontologies en RDF Schema (RDFS) et OWL seront également abordés. De plus, un exercice pratique permettra de découvrir l'utilisation concrète d'un triplestore et les bases du langage de requêtes SPARQL.

En pratique, une compréhension générale de RDFS OWL et SPARQL devrait être suffisante pour pouvoir tirer profit du Web sémantique. L'une des séances portera sur l'utilisation de l'outil AskOmics pour intégrer des données de tableurs, les convertir automatiquement en RDF, les lier avec des connaissances sous forme d'ontologies et les interroger intuitivement par des requêtes, en laissant l'outil générer le code SPARQL et l'exécuter.

Ces notions serviront de base pour l'atelier sur le raisonnement temporel et sur les méthodes d'analyse de données.

2.2 Objectifs de formation

Cette activité permettra à une personne étudiante :

- O1. de structurer des données en RDF ;
- O2. de stocker des données RDF dans un triplestore (Apache Fuseki) ;
- O3. d'interroger les données à l'aide du langage SPARQL ;
- O4. d'exploiter des ontologies pour raisonner sur les données.

3 Références

Cette section présente les principales références documentaires utilisées pour construire l'activité et les références pour approfondir des concepts présentés.

3.1 Références essentielles

Olivier Dameron, 2024. RDF-SPARQL cheatsheet
<https://gitlab.com/odameron/rdf-sparql-cheatsheet>

Bob DuCharme, 2021. What is RDF? What can this simple standardized model do for you?
<https://www.bobdc.com/blog/whatisrdf/>

Bob DuCharme, 2015. SPARQL in 11 minutes. <https://www.youtube.com/watch?v=FvGndkpa4K0>

Apache Jena, *Apache Jena Fuseki* (2024) (release 5.0.0) [logiciel], Apache,
<https://jena.apache.org/download/>

AskOmics (release 4.5.0) [logiciel], <https://askomics.org/>

3.2 Références complémentaires

Dameron, Olivier. Méthodes du Web sémantique pour l'intégration de données en sciences de la vie. Intégration de données biologiques, ISTE Group, pp.1-30, 2022, 9781789480306. ⟨hal-03720874⟩

W3C RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, 25 February 2014.
<https://www.w3.org/TR/rdf11-concepts/>

W3C SPARQL 1.1 Query Language. W3C Recommendation, 21 March 2013.
<https://www.w3.org/TR/sparql11-query/>

DuCharme, Bob. Learning SPARQL: querying and updating with SPARQL 1.1. Second edition. Sebastopol, CA: O'Reilly Media, 2013. <http://www.learningsparql.com/>