Année universitaire 2020-2021

Stage M2 en informatique ou dernière année d'école d'ingénieur avec une spécialité de développement informatique internet et systèmes d'information distribués

Titre du sujet

Plateforme internet d'accès aux données pour les Observatoires Hommes-Milieux du CNRS-INEE.

Plateforme internet d'accès aux ressources pour la science ouverte pour les 13 Observatoires Hommes-Milieux du LabEx DRIIHM (Dispositif de Recherche Interdisciplinaire sur les Interactions Hommes-Milieux du CNRS-INEE)

Mots clés

Portail web, web sémantique, web service, Web API, REST, javascript, HTML5, e-infrastructure, systèmes distribués, intergiciels, open platform, open data

Durée, rémunération et accueil

- 5 à 6 mois (date de début à partir de janvier 2021)
- gratification 591,51€/mois brut non imposable sans charges
- lieu d'accueil principal à l'IRIT : 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9 https://www.irit.fr/informations-pratiques/acceder-a-lirit/

Contact et encadrement

Si vous êtes intéressé(e), veuillez envoyer votre candidature (CV, lettre de motivation, relevés de notes si disponibles) à :

- DAYRE Pascal - Unité Mixte de Recherche IRIT/CNRS - pascal.dayre@irit.fr et
- LERIGOLEUR Emilie - UMR GEODE/CNRS - emilie.lerigoleur@univ-tlse2.fr

Nature du projet

Conception et développement logiciel d'une plateforme internet composée d'un portail web et d'une couche d'intermédiation à des web services afin d'enrichir l'éco-système numérique pour la science ouverte et l'environnement virtuel de recherche d'une communauté de recherche.

Description du contexte du sujet

Contexte de l'open research data

Faisant suite à l'open data, le mouvement de la science ouverte se structure actuellement en France à travers le Plan national pour la science ouverte (juillet 2018) visant notamment à "*Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics (sauf exceptions)*".
Le Centre national de la recherche scientifique (CNRS) a également désormais une feuille de route sur la science ouverte (2019) qui prône :

- 100% des publications scientifiques en accès ouvert (sur l'archive HAL) ;
- Développer une culture de la gestion/partage des données chez tous les acteurs du cycle de vie de la donnée (...) basée sur la mise en œuvre des principes FAIR.

Les principes FAIR fournissent des lignes directrices pour améliorer la facilité le repérage [F], l'accessibilité [A], l'interopérabilité [I] et la réutilisation [R] des ressources numériques scientifiques.
Pour ce faire, il est nécessaire de mettre en place des éco-systèmes numériques intégrant des ressources et des services en nombre toujours plus grand et facilitant la "FAIRisation" des données de la recherche.

Contexte métier

Le dispositif de recherche (LabEx DRIIHM) rassemble 13 observatoires scientifiques, les observatoires « hommes-milieux » (OHM) en France et à l'international. Depuis plusieurs années, les scientifiques impliqués étudient les dynamiques environnementales, culturelles et sociétales sur ces territoires, à différentes échelles spatiales et temporelles.

Des données hétérogènes sont produites chaque année par les équipes de recherche

multidisciplinaires (>100 projets annuels lauréats). Elles font essentiellement partie de la “longue traîne des données”, elles ne disposent donc pas d’infrastructure facilitant leur gestion et leur partage. Au-delà de leurs incontournables stockage et catalogage, il est crucial d’améliorer le partage et l’ouverture de ces données pour favoriser les analyses croisées interdisciplinaires et communiquer les résultats auprès de la société.

Le projet ANR SO-DRIIHM (2020-2023) offre un nouvel élan visant à :

1. Informer la communauté scientifique des bénéfices du partage et de l’ouverture des données
2. Créer un portail web qui facilite l'accès aux ressources existantes et la démarche de diffusion de la donnée.

Le stage s’inscrit dans le second objectif du projet SO-DRIIHM avec la création d’un premier prototype DATA-DRIIHM.

Description du travail demandé

La science ouverte s'appuie sur les principes FAIR (*Findable Accessible Interoperable Reusable*). Ces principes doivent être mis en œuvre dans un environnement facilitant la recherche, l'accès, l'utilisation et la gestion des ressources numériques afférentes.

Il s'agit de concevoir et de développer un environnement « virtuel » de travail composé de :
1/ le portail web pour la mise en œuvre des cas d'usage des principes FAIR (frontend).
2/ Web APIs pour implémenter une couche intergicielle de service (middleware) permettant l’intégration et l'accès aux différents services existants de l'éco-système numérique de la recherche aux niveaux national et international comme les services institutionnels ou les services des e-infrastructures de recherche pour alimenter le portail web (backend).

Au cours de ce stage, nous nous concentrerons principalement sur la mise en œuvre des fonctionnalités de "facilité le repérage" [F - Findable] et l'accessibilité [A - Accessible] des données selon les principes FAIR. Il s'agit de faciliter la découverte, l'exploration et l'accès des données disponibles sur les zones géographiques d'emprise des observatoires. Il sera demandé de contribuer à la spécification, de concevoir et de développer cet environnement pour les données des différentes disciplines scientifiques des observatoires. Nous parlons ici de données ou de ressources numériques au sens large.

L'IHM devra permettre les cas d'usage suivants (organisés autour du principe [F] de FAIR) sur ordinateur et sur smartphone:
[F]> Enregistrer un annuaire de référentiels de données comme re3data.org
[F]> Trouver des entrepôts de données selon une thématique scientifique
[F]> Trouver des jeux de données dans des référentiels tiers et dans un référentiel pour la

longue traîne de données sur les emprises géographiques des 13 observatoires
[F]> Trouver des publications dans des référentiels tiers
[A]> Accéder par téléchargement et par visualiser en ligne d'un jeu de données sélectionné
[A]> Accéder par téléchargement et par visualiser en ligne d'une publication sélectionnée

Le portail web offrira à l'utilisateur une IHM et un espace personnel de travail permettant d'afficher et de sélectionner la liste des référentiels de données, des disciplines et de lancer des recherches en paramétrant le service de recherche par la sélection de son observatoire OHM et par la sélection des disciplines scientifiques, de mots-clés et une emprise spatiale et/ou temporelle.
L'utilisateur pourra alors raffiner sa requête puis télécharger, visualiser les données, jeux de données par jeux de données, ou faire une visualisation croisée sur l'e-infrastructure DATA-DRIIHM.
L'utilisateur pourra réutiliser son historique de recherche.

La couche de service web devra permettre de :
- construire un registre des services de catalogage de données,
- paramétrer et enregistrer les requêtes fédérées pour récupérer la liste des méta-données et les accès aux données pour chacun des services de référentiels de données ou de catalogues de données,
- Exécuter les requêtes fédérées et restituer comme résultat les méta-données, un lien de téléchargement, un lien de visualisation de chaque jeu de données et un lien source de données pour permettre la visualisation croisée multi-sources dans l'e-infrastructure DATA-DRIIHM.

Une démonstration sur l'exploration, l'utilisation des données de la recherche, de leurs services associés, des publications dans le cadre du LabEx DRIIHM sera un attendu. Nous nous intéressons à des scénarios d'usage faisant appels à des données d'observatoire et de la longue traîne des sciences de l'univers, sciences de l'environnement et sciences humaines et sociales.

Les développements se feront selon l'état de l'art des développements internet comme par exemple les web API, les spécifications d'Open API (https://www.openapis.org/), REST pour le backend et HTML5 et les frameworks javascript pour le frontend compatible ordinateur et smartphone. Nous utiliserons les technologies du web sémantique et du W3C pour descrire et utiliser les ressources (cf. SPARQL, RDF, RDFS, OWL, DCAT https://www.w3.org/TR/vocab-dcat-3/)

Les livrables suivants sont attendus :

- les spécifications de l'IHM (frontend) et de la couche de service (backend)
- la conception générale et détaillée
- le code et la documentation détaillée

- les tests et le scénario de livraison
- le manuel utilisateur
- le manuel administrateur de la couche de service
- le manuel de déploiement
- Une machine virtuelle pour l'environnement de développement
- Une machine virtuelle et/ou des conteneurs Docker pour le déploiement

Environnement technologique

- Le cadre de Description des Ressources du W3C : JSON, Web sémantique, Sparql, RDF, RDFS, OWL, DCAT https://www.w3.org/TR/vocab-dcat-3/, ...
- Modélisation du logiciel UML
- Architectures orientées services SOA voir ROA (micro-services), REST
- Frontend : HTML5, javascript, frameworks.
- Backend : web API, web services, SPARQL endpoints (https://www.openapis.org/)
- Les ressources (W3C) : JSON, Web sémantique (RDF, RDFS, OWL, DCAT https://www.w3.org/TR/vocab-dcat-3/). Sparql
- Pour le déploiement, les conteneurs Dockers seront privilégiés afin de pouvoir faire un déploiement dans le CLOUD

Formation attendue

Masters 2 et écoles d'ingénieur de spécialité en développement logiciel des systèmes distribués sur internet (connaissance des architectures distribuées et des technologies du net).

Pour en savoir plus

SO-DRIIHM - ses premières références : (également disponibles dans HAL)
Le projet SO-DRIIHM : https://zenodo.org/record/3503385#.XkQGfGhKiUk
Communication orale : https://zenodo.org/record/3548214#.XkQFIWhKiUk
Poster : https://zenodo.org/record/3504565#.XkQGuGhKiUk

Les textes de référence clés au niveau national :

- Plan National pour la Science ouverte (4 juillet 2018): bit.ly/plannational
- Stratégie nationale des infrastructures de recherche 2018

- Feuille de route du CNRS pour la science ouverte (19 novembre 2019)

Principes FAIR :

- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- https://www.go-fair.org/fair-principles/
- FAIR is not equal to Open

Proposition de sujet de stage de M2 en informatique - Année universitaire 2020-2021

Titre du sujet

Plateforme internet distribuée d'accès aux ressources pour la science ouverte pour les 13 Observatoires Hommes-Milieux du LabEx DRIIHM (Dispositif de Recherche Interdisciplinaire sur les Interactions Hommes-Milieux du CNRS-INEE)

Mots clés

systèmes distribués, intergiciels, IHM, internet, e-infrastructure, données de la recherche, open data

Durée, rémunération et accueil

- 5 à 6 mois (date de début adaptable selon la formation)
- gratification 591,51€/mois brut non imposable sans charges
- lieu d'accueil principal à l'IRIT : 2, rue Camichel, BP 7122, 31071 Toulouse Cedex 7 ou 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9 https://www.irit.fr/informations-pratiques/acceder-a-lirit/

Contact et encadrement

Si vous êtes intéressé(e), veuillez envoyer votre candidature (CV, lettre de motivation, relevés de notes si disponibles) à :

- DAYRE Pascal - Unité Mixte de Recherche IRIT/CNRS - pascal.dayre@irit.fr et
- LERIGOLEUR Emilie - UMR GEODE/CNRS - emilie.lerigoleur@univ-tlse2.fr

Nature du projet

Conception et développement logiciel d'un système d'information distribué sur internet.

Description du contexte du sujet

Contexte de l'open research data

Faisant suite à l'open data, le mouvement de la science ouverte se structure actuellement

en France à travers le Plan national pour la science ouverte (juillet 2018) visant notamment à "*Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics (sauf exceptions)*".

Le Centre national de la recherche scientifique (CNRS) a également désormais une feuille de route sur la science ouverte du CNRS (2019) qui prône :

- 100% des publications scientifiques en accès ouvert (sur l'archive HAL) ;
- Développer une culture de la gestion/partage des données chez tous les acteurs du cycle de vie de la donnée (...) basée sur la mise en œuvre des principes FAIR.

Les principes FAIR fournissent des lignes directrices pour améliorer la facilité de repérage [F], l'accessibilité [A], l'interopérabilité [I] et la réutilisation des ressources numériques [R]. Pour ce faire, il est nécessaire de mettre en place des éco-systèmes numériques intégrant des ressources et des services en nombre toujours plus grand et facilitant la "FAIRisation" des données de la recherche.

Contexte métier

Le dispositif de recherche (LabEx DRIIHM) rassemble 13 observatoires scientifiques, les observatoires « hommes-milieux » (OHM) en France et à l'international. Depuis plusieurs années, les scientifiques impliqués étudient les dynamiques environnementales, culturelles et sociétales sur ces territoires, à différentes échelles spatiales et temporelles.

Des données hétérogènes sont produites chaque année par les équipes de recherche multidisciplinaires (>100 projets annuels lauréats). Elles font essentiellement partie de la "longue traîne des données", elles ne disposent donc pas d'infrastructure facilitant leur gestion et leur partage. Au-delà de leurs incontournables stockage et catalogage, il est crucial d'améliorer le partage et l'ouverture de ces données pour favoriser les analyses croisées interdisciplinaires et communiquer les résultats auprès de la société.

Le projet ANR SO-DRIIHM (2020-2023) offre un nouvel élan visant à :

1. Informer la communauté scientifique des bénéfices du partage et de l'ouverture des données pour tendre vers un changement progressif des pratiques de gestion et de diffusion de la donnée ;
2. Créer un portail web qui facilite la démarche de diffusion de la donnée. L'originalité de cette approche : la co-construction avec les chercheurs producteurs et (ré-)utilisateurs des données. Des spécialistes en ergonomie veilleront au design adapté du portail web.

L'approche par cycles de développement itératifs permettra d'affiner les fonctionnalités priorisées par la communauté scientifique. On peut citer par exemple pour chaque jeu de données l'attribution d'un DOI, sa visualisation (web SIG, data viz), sa diffusion via un entrepôt ou un data paper, etc., tout en veillant à l'interopérabilité des outils avec les infrastructures de données de la recherche (inter)nationales.

Le stage s'inscrit dans le second objectif du projet SO-DRIIHM avec la création d'un premier prototype DATA-DRIIHM.

Description du travail demandé

La science ouverte s'appuie sur les principes FAIR (*Findable Accessible Interoperable Reusable*). Ces principes doivent être mis en œuvre dans un environnement facilitant la gestion, l'accès et l'utilisation des ressources numériques afférentes.

Il s'agit de concevoir et de développer un environnement « virtuel » de travail composé de :
1/ une couche intergicielle permettant l'intégration et l'accès aux différents services existants de l'éco-système numérique de la recherche au niveau national et international comme les services institutionnels ou les services des e-infrastructures de recherche.
2/ une IHM web pour la mise en œuvre des cas d'usage des principes FAIR (décrit ci-après).

Une démonstration sur l'exploration, l'utilisation des données de la recherche, de leurs services associés, des publications dans le cadre du LabEx DRIIHM sera un attendu.

L'IHM doit permettre les cas d'usage suivants (organisés autour des principes F.A.I.R. [1]):
F> Enregistrer un annuaire de référentiel de données comme r3data.org
F> Trouver des entrepôts de données selon une thématique scientifique
F> Trouver des jeux de données dans des référentiels tiers et dans un référentiel local (un référentiel local sera fourni)
F> Trouver des publications dans des référentiels tiers et dans un référentiel local
A> Accéder par téléchargement ou visualiser en ligne un jeu de données sélectionné
A> Accéder par téléchargement ou visualiser en ligne une publication sélectionnée
I> Annoter un jeu de données ou une publication avec des métadonnées
I> Déposer un jeu de données avec des métadonnées sémantiques (un outil de saisie de métadonnées sera intégré)
R> Charger un jeu de données, un code et exécuter un traitement sur le jeu de données dans un datalake (une API d'un datalake en développement à l'IRIT sera utilisée)

Un service d'authentification sera développé à partir de la fédération d'identité de Renater.

Nous pouvons noter un ensemble de services à intégrer dans l'intergiciel DATA-DRIIHM par la découverte ou configuration des :

- annuaires/registres de référentiels
- registres de type de données
- référentiels de données (INIST.opidor, …)
- référentiels d'ontologies
- référentiels de méta-données
- référentiels bibliographiques et de leurs services de recherche associés (HAL, INIST.ISTEX, ...)
- référentiels de services

Il est demandé un travail de modélisation de l'IHM et de la couche intergicielle avec UML.
L'intergiciel sera modélisé avec UML en se basant sur les patrons de conception de l'orienté objet de référence pour définir des services génériques.

Il sera conçu comme un cadre applicatif générique et une couche de service intégrative proposant un ensemble de service à l'API web cliente.

Son développement se fera pour le rendre le plus générique possible à partir des

spécifications d'Open API (https://www.openapis.org/).

Les livrables suivants sont attendus :

- les spécifications de l'IHM et de la couche de service d'intermédiation (UML) générique et de l'implémentation de quelques services référents
- la conception générale et détaillée (UML)
- le code et la documentation détaillée
- les tests et le scénario de livraison (UML)
- le manuel utilisateur
- le manuel de déploiement
- le manuel administrateur de la couche de service (gestion des services, ajout d'un registre, d'un référentiel, gestion des utilisateurs...)
- Une machine virtuelle pour l'environnement de développement
- Une machine virtuelle et des conteneurs Docker pour le déploiement

Environnement technologique

- Pour la modélisation et la conception : UML

- Pour le développement :

Les technologies du WEB (W3C, javascript, web sémantique, navigateurs,...) et des architectures orientées services SOA voir ROA (micro-services), openAPI (https://www.openapis.org/).

- Pour le déploiement, les conteneurs Dockers seront privilégiés afin de pouvoir faire un déploiement dans le CLOUD.

Formation attendue

Développement Logiciel des systèmes distribués (connaissance des architectures distribuées et des technologies du net)

Pour en savoir plus

SO-DRIIHM - ses premières références : (également disponibles dans HAL)
Le projet SO-DRIIHM : https://zenodo.org/record/3503385#.XkQGfGhKiUk
Communication orale : https://zenodo.org/record/3548214#.XkQFIWhKiUk
Poster : https://zenodo.org/record/3504565#.XkQGuGhKiUk

Les textes de référence clés : (liste non exhaustive)
Au niveau européen ou paneuropéen :

- DIRECTIVE (UE) 2019/1024 DU PARLEMENT EUROPÉEN ET DU CONSEIL du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public
- EU Commission, 2016, H2020 Programme Guidelines on FAIR Data Management in Horizon 2020
- Turning FAIR into reality, Final report and action plan from the European Commission expert group on FAIR data, European Commission (2018): bit.ly/turningFAIR
- Goudeseune Lise, et al.. (2019, September 19). Guidance document for scientists on data management, open data, and the production of Data Management Plans. BiodivERsA report.

Zenodo. http://doi.org/10.5281/zenodo.3448251

Au niveau national :

- Plan National pour la Science ouverte (4 juillet 2018): bit.ly/plannational
- Stratégie nationale des infrastructures de recherche 2018
- Feuille de route du CNRS pour la science ouverte (19 novembre 2019)

Principes FAIR :

- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- https://www.go-fair.org/fair-principles/
- FAIR is not equal to Open