Le service de gestion de données FG-iRODS : une infrastructure fédérée pour l'hébergement de données scientifiques à l'échelle nationale et européenne.

Catherine Biscarat
Laboratoire de Physique Subatomique et de Cosmologie
53, avenue des Martyrs
38026 Grenoble

Raphaël Flores
URGI, INRA, Université Paris-Saclay
Unité de Recherche Génomique-Info
Route de Saint-Cyr
78026 Versailles Cedex

Pierre Gay
Mésocentre de Calcul Intensif Aquitain
351, cours de la Libération CS 10004
33 405 Talence CEDEX

Christine Gondrand
Laboratoire de Physique Subatomique et de Cosmologie
53, avenue des Martyrs
38026 Grenoble

Emmanuel Medernach
Institut Pluridisciplinaire Hubert Curien
23, rue du Loess – BP28
67037 Strasbourg Cedex 2

Patrick Moreau
Laboratoire Agroécologies, Innovations et Ruralités
24, chemin de Borde-Rouge
31326 Auzeville

Vincent Negre
Laboratoire d'Écophysiologie des Plantes sous Stress Environnementaux
2, place Viala
34060 Montpellier Cedex 2

Jérôme Pansanel
Institut Pluridisciplinaire Hubert Curien
23, rue du Loess – BP28
67037 Strasbourg Cedex 2

Geneviève Romier
Centre de Calcul de l'IN2P3
21, avenue Pierre de Coubertin CS70202
69627 Villeurbanne Cedex

Résumé

Depuis plusieurs années, l'ensemble des disciplines scientifiques font face à un déluge de données hétéroclites, tant par les moyens d'acquisition (collecte de terrain, expérimentation, modélisation, simulation, etc.), que par les formats. Fournir des ressources de stockage en quantité suffisante aux chercheurs ne suffit plus pour répondre aux nouvelles questions scientifiques. Il faut être en mesure de les exploiter aujourd'hui, tout comme dans le futur. La conservation, la description et le partage des données doivent donc être pensés dès le début des projets pour répondre aux principes du FAIR.

Afin d'accompagner et de répondre aux besoins des chercheurs sur ces thématiques, plusieurs laboratoires partenaires de France Grilles ont développé depuis 2013 une expertise sur le logiciel iRODS et mis en place une infrastructure fédérée s'appuyant sur des ressources géographiquement distribuées pour fournir un service de gestion de données, dénommé FG-iRODS.

Ce service mutualisé permet, par exemple, de faciliter la gestion de grandes collections de données et les métadonnées associées, de gérer finement les droits d'accès aux fichiers et d'automatiser les flux de données.

Cet article détaille les objectifs du projet FG-iRODS, la nouvelle infrastructure matérielle et logicielle, le service proposé et un cas concret d'utilisation.

Mots-clefs

iRODS, gestion de données, données scientifiques, stockage, mutualisation d'expertise, service, infrastructure distribuée, France Grilles

1 Introduction

France Grilles, groupement d'intérêt scientifique (GIS) créé en 2010, a pour missions d'établir, d'opérer et de fédérer des moyens de calcul et de stockage géographiquement distribués au bénéfice de la recherche scientifique en France. En 2013, le service FG-iRODS a été mis en place afin de répondre à la demande des utilisateurs souhaitant disposer d'un service de gestion de données accessibles à la fois depuis la grille de calcul, depuis des serveurs hébergés dans le service de « *Cloud Computing* » ou depuis leur ordinateur personnel.

Le déploiement de ce nouveau service a fédéré plusieurs laboratoires souhaitant à la fois partager leur expertise autour du logiciel iRODS et mettre en commun des ressources de stockage autour d'un catalogue centralisé. Chaque partenaire participe au pilotage du projet.

Le service héberge actuellement plusieurs projets scientifiques et l'infrastructure sous-jacente a été renouvelée en 2018.

2 FG-iRODS

2.1 Stratégie

Le projet FG-iRODS vise au développement et à la mise en production d'un service de gestion de données destiné à répondre aux besoins des chercheurs. L'accès aux ressources doit être ouvert à toutes les communautés scientifiques et être interopérable avec les autres infrastructures de stockage de données couramment déployées. Ce service se positionne donc comme complémentaire des autres acteurs du domaine, et compte tenu des ressources actuelles, est plutôt destiné au projet de petites et moyennes envergures. À partir de cette stratégie, les objectifs suivants ont été fixés :

- développer une communauté d'experts autour d'iRODS et de la gestion des données scientifiques ;
- valoriser et partager leur expérience par le biais de publications et d'interventions lors de formations et de conférences ;
- créer une infrastructure de niveau production sur laquelle est déployé le logiciel iRODS ;
- disposer d'une offre de service détaillée et d'une procédure d'accueil des utilisateurs.

2.2 Pilotage

Le projet est piloté par un groupe composé d'administrateurs système, de développeurs et d'utilisateurs. Ce groupe se réunit régulièrement par visioconférences, au cours desquelles le point est fait sur le fonctionnement des sites et du service, le suivi des utilisateurs, les développements en cours ainsi que la programmation des formations.

Ces réunions permettent de suivre l'évolution de l'infrastructure, d'échanger autour des problèmes d'administration des services et des opérations, et de maintenir l'adéquation du service au besoin des utilisateurs.

Des outils collaboratifs ont été mis en place pour mutualiser au mieux expériences et bonnes pratiques (wiki, listes de discussion, hébergement de code, partage de supports de formation).

3 Le service FG-iRODS

3.1 Infrastructure

Le service repose sur une fédération de ressources de stockage réalisée à l'aide du logiciel iRODS, actuellement en version 4.2.

Ce logiciel développé et maintenu par un large consortium, est un logiciel libre disponible pour les différentes distributions GNU/Linux. Il permet d'accéder, de gérer et de partager des données stockées sur différents types de stockage et facilite ainsi l'accès à des ressources hétérogènes (Unix, stockage objet S3, stockage sur bande HPSS, etc.). Sa puissance est liée à son moteur de règles et aux micro-services, qui offrent la possibilité de contrôler finement les données. Cette solution rend possible :

- la virtualisation de l'accès au stockage ;
- la gestion de plusieurs péta-octets de données ;
- la parallélisation des transferts de données volumineuses ;

– la gestion des métadonnées ;

– l'automatisation des processus ;

– la sécurisation des données grâce à la réplication et à la gestion fine des accès.

L'infrastructure regroupe actuellement sept sites répartis en France (Figure 1) et permet aux utilisateurs d'accéder de façon transparente à 1,3 péta-octets de stockage. L'ensemble de l'infrastructure a été renouvelée en 2018.

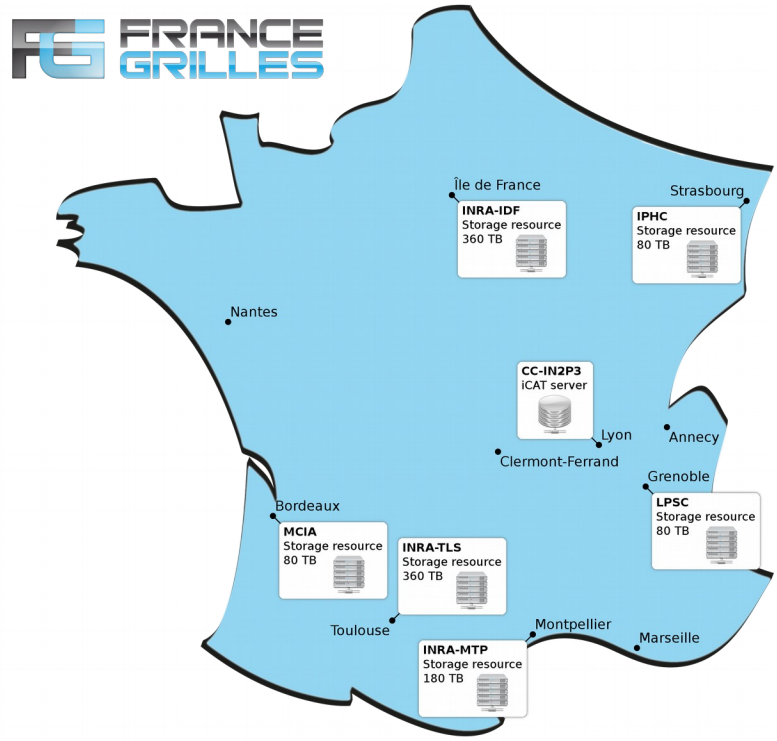

Figure 1 - Carte de France des sites participant à l'infrastructure iRODS

3.2 Extension de l'infrastructure

Le projet est ouvert aux nouveaux partenaires souhaitant intégrer à l'infrastructure FG-iRODS leurs ressources de stockage propres afin de bénéficier de ce réseau d'expertise, ainsi que d'une administration mutualisée du catalogue iRODS. Les demandes sont étudiées spécifiquement, les points suivants étant étudiés :

– impact sur le service de l'ajout des nouvelles ressources ;

– contribution du nouveau partenaire à la co-administration du service ;

– compatibilité de l'infrastructure à connecter.

3.3 Développements réalisés

Afin de faciliter l'exploitation de l'infrastructure et son utilisation, un ensemble d'outils ont été développés et / ou déployés :

- outil de surveillance fonctionnel basé sur Nagios ;
- comptabilité des ressources utilisées par les utilisateurs France Grilles ;
- interfaces graphiques (Brocoli et MetalNX) ;
- chiffrage des échanges par SSL.

Des développements sont en cours pour intégrer l'authentification OpenID.

3.4 Services pour les utilisateurs

Afin de faciliter l'utilisation de FG-iRODS par les utilisateurs, l'offre comporte plusieurs services. Cette liste n'est pas figée et peut être étendue en fonction des nouvelles demandes.

3.4.1 Formation et documentation

Des formations sont organisées à la demande pour les utilisateurs de FG-iRODS. Elles peuvent être couplées avec d'autres formations, telle que l'utilisation de DIRAC. La formation iRODS utilisateur dure deux jours, et peut être complétée par une journée sur le développement de règles.

Ces formations sont complétées par une documentation complète disponible en ligne, ainsi que par les conseils du réseau d'experts.

3.4.2 Accès depuis les infrastructures de calcul

Les données stockées sur l'infrastructure iRODS peuvent être utilisées à la fois sur l'infrastructure de Cloud Computing (authentification par mot de passe) ou par la grille de calcul (authentification par certificat). Plusieurs centres nationaux et mésocentres (IDRIS, GRICAD, etc.) ont également déployé les clients iRODS sur leurs systèmes, rendant les données accessibles depuis ces infrastructures.

3.4.3 Développement de règles

En fonction des besoins des utilisateurs, des règles peuvent être développées pour automatiser la gestion des données. Les utilisateurs peuvent également proposer leurs propres règles, mais elles ne seront intégrées qu'après validation par les administrateurs du service. Le moteur de règle iRODS permet par exemple :

- de répliquer automatiquement certaines données ;
- d'extraire des méta-données en analysant les fichiers ;
- de lancer automatiquement des traitements lorsque les données sont copiées ;
- de convertir les données.

4 Utilisation du service

4.1 Accès à FG-iRODS

Le service FG-iRODS, tout comme les autres services France Grilles, est accessible pour la réalisation des activités scientifiques de l'ensemble des partenaires du GIS, ainsi qu'aux organismes et entreprises ayant des projets communs avec eux.

Afin de pouvoir utiliser le service FG-iRODS, il faut contacter France Grilles à travers l'adresse de contact disponible sur la page Web du service FG-iRODS. En retour, un formulaire d'accueil est envoyé, permettant au futur utilisateur de détailler sa demande. La demande est ensuite étudiée en réunion par les membres du comité de pilotage du service. Il est à noter que l'accès au service iRODS est assujetti à la fourniture d'un plan de gestion de données.

L'utilisation de FG-iRODS n'est pas facturée. En effet, les infrastructures ont été financées par France Grilles. Toutefois, si les besoins de l'utilisateur sont trop importants par rapport à la taille de l'infrastructure, un achat d'équipement (mutualisation) pourra être demandé.

Afin de justifier la pertinence scientifique de son infrastructure et de ses services, France Grilles demande aux utilisateurs :

- d'ajouter une mention à France Grilles dans la section « remerciements » de leurs publications issues de l'utilisation de ce service ;
- de référencer ces publications dans HAL afin de pouvoir les ajouter à la collection France Grilles.

Les utilisateurs sont également invités à présenter leurs travaux lors de conférences (JCAD, etc.).

4.2 Exemple d'utilisation

Le projet Phenome-Emphasis, associant l'INRA, Arvalis et Terres-Inovia ambitionne de développer des infrastructures de phénotypage haut-débit au niveau national. Les systèmes d'acquisition aux champs (drone, phenomobile) embarquent différents capteurs (caméras haute résolution RGB, multispectrales et infra-rouge thermique, LIDARs) qui génèrent un volume important d'images qu'il convient de traiter, stocker et archiver. Ces données sont analysées à travers un flux de traitements automatisé basé sur l'enchaînement de services en conteneur.

Les données brutes et analysées sont stockées sur l'infrastructure FG-iRODS et représenteront à l'issue du projet une volumétrie d'environ un péta-octet. Les ingénieurs de ce projet participent à l'administration du service FG-iRODS. Compte tenu de la volumétrie produite, le projet Phenome-Emphasis a financé l'achat des ressources de stockage nécessaires (serveurs DELL R730xd couplés à des baies de stockage MD1400). Ces ressources de stockage sont hébergées dans trois datacentres (INRA à Toulouse, CEA en Île-de-France et CINES à Montpellier) et intégrées à l'infrastructure FG-iRODS : le catalogue est celui du service FG-iRODS, mutualisé au centre de calcul de l'IN2P3.

5 Conclusion

Le service FG-iRODS est en production depuis maintenant cinq ans. L'année 2018 a été une année charnière, avec une évolution majeure sur plusieurs points :

- accueil du projet Phenome-Emphasis ;
- renouvellement de l'infrastructure (ressources de stockage et catalogue iRODS) ;
- migration vers iRODS v4 ;
- évolution de la procédure d'accueil des utilisateurs.

Ces évolutions ont permis au projet de franchir une nouvelle étape et de répondre à des demandes plus importantes. Des travaux complémentaires sont en cours afin d'améliorer l'intégration des outils avec les plateformes existantes, notamment en termes d'interopérabilité, avec la mise en place de l'authentification OpenID ou d'une interface WebDAV.