AgroDataRing. Une infrastructure partagée et mutualisée pour le stockage longue durée

Pierre Adenot, Stéphane Bansard, David Benaben, Véronique Brunaud, Christophe Caron, Alexandre Dehne-Garcia, Christophe Duperier, Adrien Falce, Olivier Filangi, Franck Giacomoni, et al.

AgroDataRing. Une infrastructure partagée et mutualisée pour le stockage longue durée

Pierre Adenot[1], Stéphane Bansard[2], David Benaben[3], Véronique Brunaud[4], Christophe Caron[5], Alexandre Dehne-Garcia[6], Christophe Duperier[7], Adrien Falce[8], Olivier Filangi[9], Franck Giacomoni[7*], Fabienne Granier[10], Philippe Grevet[4], Nicolas Guilhot[11], Annie Hofstetter[12], Johann Joets[8], Thierry Hotelier[13], Olivier Langella[8], Ludovic Legrand[14], Mikael Loaëc[15], Virginie Lollier[2], Patrick Moreau[16], Emmanuelle Morin[17], Hadi Quesneville[15], Tovo Rabemanantsoa[18], Gérald Salin[19], Dominique Tessier[2]

Résumé. Les communautés scientifiques se trouvent aujourd'hui confrontées à un changement de paradigme autour de la gestion des données, nécessitant une meilleure gestion du cycle de vie des données avec notamment leur traitement et intégration, et leur partage. À la suite du chantier « Data Partage » lancé dès 2012 à l'Inra, le groupe de travail « e-infra Storage » a initié en 2016 une réflexion collective autour des besoins de l'Institut en matière de dispositif de stockage des données patrimoniales scientifiques qui a abouti à la co-construction d'une infrastructure partagée et mutualisée : AgroDataRing.

Mots clés : données, stockage, pérennisation, mutualisation, partage, communautés

Remerciements : Mathieu Andro, Philippe Cheneau, Eddie Iannuccelli, Thomas Lallart, Lionel Million, Sylvie Nugier, Stéphane Paris, Rémi Person, Lina Sbeih, Florian Trincal, Magalie Weber

[1] UMR BDR Biologie du Développement et Reproduction, Inra, Enva, Université Paris Saclay, 78350 Jouy en Josas, France
[2] UR BIA Biopolymères, Interactions Assemblages, Nantes, France
[3] UMR BFP Biologie du Fruit et Pathologie, Bordeaux, France
[4] UMR IPS2 Institut des Sciences des Plantes de Paris Saclay Versailles-Grignon, France
[5] UAR IngeNum Ingénierie Numérique en Recherche, Toulouse, France
[6] UMR CBGP Centre de Biologie pour la Gestion des Populations, Montpellier, France
[7] UMR UNH Unité de Nutrition Humaine, Clermont-Ferrand, France
[8] UMR GQE Génétique Quantitative et Evolution, Le Moulon, France
[9] UMR IGEPP Institut de Génétique Environnement et Protection des Plantes, Rennes, France
[10] UMR IJPB Institut Jean-Pierre Bourgin Versailles-Grignon, France
[11] UMR GDEC Génétique Diversité et Ecophysiologie des Céréales, Clermont-Ferrand, France
[12] UMR LAMETA Laboratoire Montpelliérain d'Économie Théorique et Appliquée, Montpellier, France
[13] Montpellier SupAgro, France
[14] UMR LIPM Laboratoire des Interactions Plantes Micro-organismes, Toulouse, France
[15] UR URGI Unité de Recherche Génomique-Info, Versailles-Grignon, France
[16] UMR AGIR AGroécologie, Innovations, teRritoires, Toulouse, France
[17] UMR IaM Interactions Arbres/Micro-organismes, Nancy, France
[18] UMR ISPA Interaction Sol Plante Atmosphère, Bordeaux, France
[19] US GeT-PlaGe Génome et Transcriptome - Plateforme Génomique, Toulouse, France
*Auteur de correspondance franck.giacomoni@inra.fr

Introduction

Les communautés scientifiques se trouvent aujourd'hui confrontées à un changement de paradigme autour de la gestion des données. Ces évolutions nécessitent une meilleure gestion du cycle de vie des données avec notamment leur traitement et intégration, et leur partage. À l'Inra, une réflexion a déjà été menée sur la gestion et le partage de la donnée scientifique dans le cadre du projet "Data Partage", et plus largement autour des tendances OpenData.

Les plateformes, laboratoires et autres dispositifs de production de données génèrent une volumétrie croissante de données liée à l'évolution des technologies. Ces données, parfois coûteuses à obtenir, sont de plus en plus souvent partagées au sein de projets de recherche intégratifs avec de nombreux partenaires. Ces évolutions introduisent un besoin de conservation des données à moyen/long terme à un coût maîtrisé et nécessitent une meilleure gestion de leur cycle de vie. Il devenait donc indispensable de disposer d'une vision plus globale des besoins de l'Institut et de proposer une solution en termes de stockage pour les données patrimoniales scientifiques.

Cet article décrit la méthodologie et les résultats d'une étude menée en 2016 par le groupe de travail « e-infra Storage » et une instanciation en 2017 des recommandations qui en ont découlé, avec la mise en place d'un projet fédérateur pour le stockage moyen/long terme : AgroDataRing (ADR). Il présente d'abord les éléments de la phase de réflexion, puis la co-construction d'une infrastructure partagée et mutualisée, dans le cadre d'une démarche collaborative en partenariat entre plusieurs Départements.

Phase de réflexion préalable

En 2016, les objectifs du groupe de travail « e-infra Storage » étaient de réfléchir à de nouveaux modes d'organisation et de proposer des solutions technologiques autour du stockage adaptées aux besoins des communautés. Cette étude a reposé sur une identification non exhaustive, mais représentative des besoins au travers des retours de différents types d'organisation générant des données à caractère scientifique : laboratoires, plateformes de production, équipes de recherche, infrastructures nationales, métaprogrammes, etc. Nous avons notamment interviewé le Centre Informatisé du Traitement de l'Information en Sciences Economiques et Sociales (CATI CITISES), des plateformes nationales (GeT-PlaGe, MIMA2, PAPSSO…) et les dispositifs nationaux Phénome et MetaboHUB. Le panel des producteurs de données interviewés se veut relativement représentatif en terme d'organisation (CATI, Plateforme, Infrastructures Nationales …) et couvre un large spectre de communautés (SHS, Imagerie, Omiques, etc.).

Expression des besoins

La stratégie portant sur la gestion des données s'est révélée être très hétérogène d'une structure à l'autre. Cependant, le besoin de pouvoir pérenniser les données au sein d'un socle commun est apparu comme central pour l'ensemble des acteurs.

Nous avons aussi mis en évidence que la volumétrie à sécuriser parfois conséquente à l'échelle d'un producteur de données restait maîtrisable à l'échelle d'un Institut. On peut ainsi estimer la production de données scientifiques de l'ordre de quelques pétaoctets par an à l'échelle de l'Inra. Il existe aujourd'hui des solutions permettant de gérer sans difficultés majeures ces volumétries.

La question du financement pérenne est également un point récurrent évoqué qu'il fallait prendre en compte avec les modes de financement actuels de la recherche publique (CPER, mode projet, etc.). Enfin sur l'échantillon choisi, nous avons constaté que l'usage des métadonnées était relativement peu répandu et très dépendant des producteurs au regard des enjeux affichés par les projets scientifiques.

De cet état des lieux, il a paru nécessaire de disposer de solutions évolutives et modulaires afin de répondre aux variations inhérentes au monde de la recherche et aux ruptures technologiques (nouveaux systèmes d'acquisition, etc.).

Les principaux besoins exprimés en lien avec les différents états du cycle de vie de la donnée portaient principalement sur :

- le stockage et la sécurité des données à un coût maîtrisable ;
- la capacité de partager facilement les données dans le cadre de collaborations ;
- l'accès simple aux données au travers d'interfaces adaptées ;
- des mécanismes permettant une gestion des métadonnées.

Bilan

Cette étude, initiée en 2016 et conclue début 2017, a démontré qu'il existait une forte diversité des données (origine, nature, volumétrie, etc.) et de leur gestion au sein des laboratoires.

Les conclusions de cette étude ont mis l'accent sur l'intérêt d'un environnement de stockage distribué permettant de répondre aux enjeux technologiques et organisationnels avec des coûts raisonnables. Un environnement de stockage distribué a pour avantage de fédérer des communautés utilisatrices, de mobiliser des expertises présentes au sein des Unités de l'Institut, de favoriser la montée en compétences et de faciliter les partages de données. Aujourd'hui de nombreux instituts aux échelles nationales (e.g. Cirad) ou internationales se tournent vers des solutions de stockage distribuées. En déployant ce type d'infrastructure, l'Inra rejoindrait ainsi une communauté plus large en se préparant aux approches d'interopérabilité et en proposant un service adapté à ses utilisateurs au travers d'une infrastructure partagée et mutualisée.

Co-construction d'une infrastructure partagée et mutualisée : AgroDataRing

Début 2017, les besoins exprimés par un premier cercle d'Unités du Département Biologie Amélioration des Plantes (BAP) ont permis d'instancier une première solution de stockage distribué : AgroDataRing (ADR) dans le cadre d'un projet fédérateur. Les quatre premières Unités partenaires (IPS2, GQE, IJPB, GDEC) ont été rejointes durant le premier trimestre par plusieurs autres structures Inra (IGEPP, BFP, UNH, MetaboHUB, IAM, BIA) adhérant à la démarche en vue de contribuer activement à l'infrastructure. L'Unité URGI a également contribué au dispositif au travers d'une expertise dans le domaine du stockage, et sera utilisatrice de l'infrastructure dans le cadre d'un PRA (plan de reprise d'activités.

Les principaux besoins exprimés par ces structures correspondaient bien à la synthèse des besoins réalisée par l'étude 2016, à savoir de disposer à un coût raisonnable d'un espace de volumétrie conséquente (plusieurs dizaines de To) pour assurer la pérennité de données scientifiques, mais aussi leur partage, tout en étant évolutif.

Fonctionnalités

ADR est un **projet fédérateur** qui vise à déployer une solution de **stockage mutualisée des données** scientifiques (images, NGS, etc.) produites en masse par les communautés. Il repose sur un modèle novateur basé sur la **contribution et le partage de compétences**. Cette approche permet de fédérer différents acteurs autour d'une seule et même solution, tout en proposant une certaine souplesse afin de s'adapter, dans la mesure du possible, aux spécificités de chacun des contributeurs.

Le projet AgroDataRing a pour objectifs :

- de répondre aux **besoins actuels et futurs pour du stockage moyen/long terme des données scientifiques à forte volumétrie ;**
- de disposer d'un **coût de stockage raisonnable et totalement maîtrisé** de l'ordre de 30 € To/an ;
- de développer des **synergies** autour d'une question qui se veut centrale dans un cadre plus large de gestion de données (partage, etc.) ;
- d'intégrer cette infrastructure dans **un contexte local/régional (30% de l'infrastructure) et DataCenter Inra (70% de l'infrastructure)** afin de tenir compte du contexte scientifique, des problématiques techniques mais aussi de partenariats régionaux (e.g. lien Mésocentre de région) ;
- d'élargir et **partager des compétences en vue de développer les expertises existantes** notamment en terme d'environnement distribué ;
- et de favoriser les **collaborations scientifiques** au travers de cet objet intégrateur qu'est ADR.

Modèle organisationnel

La gouvernance proposée se veut légère et efficiente afin d'articuler les niveaux décisionnaires et organisationnels en précisant les rôles des différents acteurs. Elle permet ainsi de s'adapter aux évolutions rapides du monde de la recherche et de garantir une certaine pérennité de la solution.
La gouvernance s'articule autour de trois axes :

- **stratégique** : les directions des laboratoires concernés décident des moyens affectés : unité de stockage et/ou RH pour l'année n+1. Leur rôle est l'identification des besoins, l'orientation des choix d'organisation et d'évolution. Ce comité est en lien avec le coordinateur Ingenum de ce projet fédérateur ;
- **pilotage** : un comité de pilotage assure les choix techniques et a un rôle décisionnaire. Il proposera les règles au niveau stratégique : accès, adhésion, gestion des RH, usage des espaces, etc. Ce comité de pilotage se réunit tous les mois durant la mise en place de l'infrastructure, puis une fois par trimestre pour la suite du projet ;
- **technique** : le comité technique assure le côté opérationnel. Il est représenté par un collectif d'ingénieurs système qui administre et fait évoluer l'infrastructure en synergie avec les autres axes. Il fonctionne en mode agile avec des points d'étapes dont la fréquence a été, par exemple, plus importante lors du déploiement initial (1 fois par semaine).

Modalités d'adhésion

L'adhésion à l'infrastructure mutualisée repose sur des principes qui précisent les engagements :

- adhésion au modèle organisationnel basé sur une **approche coopérative** et respect des spécifications techniques ;
- apport *a minima* de deux **équipements** de stockage que l'on appelle « briques » d'une capacité unitaire de 80 To utiles ;
- fourniture et partage de l'expertise (*a minima* 0,1 ETP) nécessaire au déploiement et à la gestion de ce projet fédérateur. En l'absence de moyen humain, cette question peut se gérer au niveau du Département de rattachement du contributeur par une mutualisation entre plusieurs Unités du même Département, ce qui est le cas pour un des contributeurs actuels (Institut Jean-Pierre Bourgin / Versailles) ;
- un engagement pour une durée minimale de 5 ans.

En adhérant au projet, le contributeur devient **codétenteur de l'infrastructure**, et à ce titre **il intègre les différentes instances de gouvernance** (comité stratégique, comité de pilotage et comité technique). Il peut ainsi être force de proposition en vue de faire évoluer l'infrastructure.

Architecture technique

L'architecture physique repose sur deux composants principaux :

- la brique de niveau 1 (proximité) est un serveur de stockage (type Dell R730XD) et propose un accès simple aux communautés servies. Elle peut être localisée dans un Data Center Inra ou régional en fonction des contraintes du contributeur ;
- la brique de niveau 2 (distante) localisée dans le Data Center Inra de Toulouse se veut très évolutive au travers de technologies innovantes (e.g. Software Defined Storage) puisqu'elle contribue à un espace complètement mutualisé.

Les couches de distribution plus intégratives seront assurées par des outils standards (RSYNC), *via* des protocoles bien établis et sécurisés (e.g. S/FTP), et à terme par des applicatifs communautaires (e.g. iRODS) afin de garantir une bonne interopérabilité avec d'autres instituts (e.g.Cirad), et ainsi proposer un modèle ouvert.

L'ensemble des briques de stockage est administré par le collectif des ingénieurs système (6/7 personnes) et fait partie de la même infrastructure.

La **Figure 1** représente schématiquement l'architecture et les services d'ADR.

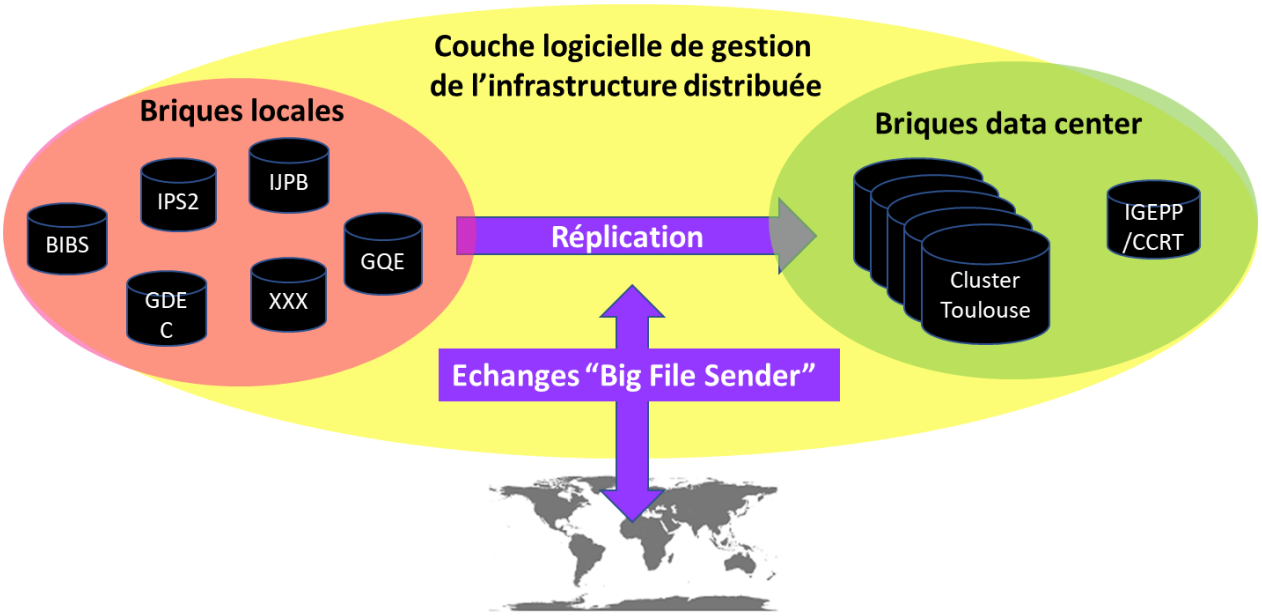

Figure 1. *Architecture et services d'ADR.*

Modèle économique

ADR n'est pas une offre de service mais un **projet fédérateur s'intégrant aux démarches collaboratives que l'on retrouve dans les environnements de recherche académiques** et en lien étroit avec les projets scientifiques. Afin de définir un modèle économique tout aussi pérenne que l'infrastructure, nous avons tenu compte du contexte, et en particulier du mode actuel de financement de la recherche. Le mode contributif nous semble opportun pour déployer et maintenir ce type de solutions. Les demandes de moyens de type équipement *via* les programmes de recherche auprès des différentes agences de moyens ou collectivités sont relativement courants.

En résumé, le ticket financier d'entrée pour un contributeur correspond à un investissement sur 5 ans de ~25 k€ et comprend :

- deux serveurs de stockage (briques pour 80 To répliqués : 160 To utiles) pour un coût total de 18 k€ avec une maintenance sur 5 ans à financer au démarrage ;
- le coût éventuel pour un hébergement sec de ces 2 serveurs : 7,5 k€ pour les tarifs Inra DataCenter (750 € / serveur / an x 5 ans). Ce coût peut varier en fonction de la localisation des serveurs et de la politique de l'hébergeur.

En tenant compte de ces éléments, le coût normalisé revient à 30 €/To/an pour les données non répliquées, et à 60 €/To/an avec une réplication. Cet indicateur (To/an) est donné à titre indicatif en lissant l'investissement et le fonctionnement sur 5 ans. Il va aussi de soi que la réplication est un mécanisme natif à la problématique ciblée par ADR.

Conclusion et perspectives

Ce projet a été construit en deux temps avec une phase d'analyse et de recueil des besoins, suivie d'une phase de co-construction. Il est également important de tenir compte de l'évolution des besoins durant la phase de fonctionnement, ce qui a été rendu possible par une architecture modulaire et un certain pragmatisme de l'ensemble des acteurs.

ADR est une solution qui se veut tout d'abord évolutive afin de garantir pour les Unités un certain niveau de pérennité. Son modèle repose aussi sur une organisation agile, totalement adaptée au monde de la recherche et aux disruptions qu'il peut connaître. En mode opérationnel depuis quelques mois, la solution s'avère être en capacité de passer à l'échelle avec l'accueil de nouveaux contributeurs tout en répondant aux besoins.

ADR s'intègre au contexte institutionnel avec une utilisation des DataCenter Inra pour les ⅔ du stockage tout en fédérant des communautés au travers d'une approche mutualisée. Cette approche a remporté un franc succès qui va bien au-delà du coût de revient. En effet le socle du projet repose sur un modèle de co-construction avec les Unités/Départements, tout en assurant une coordination nationale (Ingenum) afin de conserver un dispositif cohérent. Par ailleurs, ADR prend en compte les modalités de financement actuel de la recherche académique (CPER, sur projet, etc.) en proposant un modèle souple et adaptatif.

La version actuelle est en production et propose près de 2 Po de stockage en mode répliqué pour une dizaine de contributeurs hébergeant plus de 1500 agents. Les briques de stockage de niveau 1 sont opérationnelles depuis décembre 2017, alors que la synchronisation vers la brique mutualisée de niveau 2 a été mise en place en février 2018. Nous envisageons de proposer de nouvelles fonctionnalités, notamment autour du partage des données, mais aussi des métadonnées. Cette dynamique se traduit aussi avec l'arrivée de plusieurs nouveaux contributeurs dès 2018.

La principale difficulté en lien avec une organisation fortement mutualisée est qu'elle repose sur des composants et des compétences réparties sur de nombreux sites. Cette difficulté initiale s'est au final transformée en point positif avec notamment du partage et de la montée en compétences des acteurs du projet. Des formations par des prestataires externes ont déjà été organisées à destination de la communauté des administrateurs de l'ADR. Ces interactions ont permis ainsi de développer de fortes synergies technologiques, voire scientifiques et d'avancer rapidement.

Rassembler les forces de plusieurs Unités pour élaborer et gérer ce type d'infrastructure partagée et mutualisée permet aussi d'envisager des connexions avec d'autres Instituts. Nous avons ainsi en perspectives de nous

articuler avec certains programmes d'e-Infrastructures institutionnelles mais aussi des projets nationaux, voire européens à terme. Enfin, de nouvelles structures nationales vont intégrer dès début 2018 cette dynamique et nous permettre ainsi de continuer à proposer de nouvelles fonctionnalités pour les communautés scientifiques.

Hommage à Christophe Caron

Il y a de cela quelques années, l'annonce du retour dans l'Institut de Christophe, après quelques temps passés sous les embruns bretons, a suscité l'enthousiasme de nombre d'entre nous. D'abord à l'idée de retrouver un collègue, et souvent un ami, très apprécié. Mais aussi, parce que ses qualités, ses compétences, son audace et son énergie semblaient avoir été placées au bon endroit. De fait, la marmite à projet n'a pas tardé à bouillir et, parmi bien d'autres choses, est né l'AgroDataRing.

Sa longue pratique du métier, sa connaissance très fine de nos environnements et de nos objectifs et sa passion pour les sciences ont permis à Christophe de se forger une vision d'une rare clarté sur les orientations à prendre et les évolutions à mener pour maximiser l'apport de l'informatique dans nos programmes de recherche. Christophe était viscéralement convaincu que les défis auxquels nous sommes confrontés ne pouvaient être abordés qu'en rassemblant les hommes et les femmes de l'Institut autour de projets fédérateurs. Si l'idée n'était pas nouvelle, y parvenir n'est pas si simple, car il n'existe pas de méthode clé en main. Au-delà des qualités déjà évoquées et de la capacité à concevoir des projets pertinents et à transmettre, il y a tout un tas d'éléments qu'on ne saurait décrire, toute une alchimie, que Christophe maîtrisait et qui lui permettait de susciter l'adhésion. En quelques mois la communauté AgroDataRing s'est mise à l'ouvrage et continue à travailler presque comme s'il en avait toujours été ainsi et on a du mal à se souvenir qu'au début, bien peu y croyaient.

Christophe nous laisse aussi une partie de son énergie. Si un jour elle venait à manquer à l'un ou à l'autre, il nous suffira de nous souvenir.

Les membres de la communauté AgroDataRing.