

Journées EOSC France 2022

Table ronde 2 (5 avril): EOSC : quelles devraient être les prochaines étapes ?

Participants:

- Sabine Crépé-Renaudin (CNRS-IN2P3)
- Odile Hologne (INRAE)
- Frédéric Huynh (Data Terra)
- Volker Beckmann (MESRI)

Modération:

- Anne Vanet (Université Paris-Cité)

1 Services : quels besoins à court/moyen terme en France ? Quels besoins de développements transverses aux disciplines scientifiques ?

OH : Distinguer trois types d'utilisateurs : chercheurs, e-infrastructures (services autour des données), services d'infrastructures.

Concernant les chercheurs, l'EOSC-Portal ne s'adresse pas à un chercheur qui n'est pas déjà embarqué dans une communauté utilisatrice d'EOSC. Premier besoin : rendre des services accessibles. Certains existent (EUDAT, OPENAire), mais le puzzle a besoin d'être assemblé.

Pour les e-infrastructures : mettre en œuvre au niveau français des services qui existent déjà, par exemple AAI, intégrer les services de base de l'EOSC pour attribuer des identifiants permanents à nos données

Pour les fournisseurs de services d'infra, Data Centers, Mésocentres : pourquoi et comment peuvent-ils être connectés à EOSC (y compris questions économiques) ?

Assembler les pièces d'un puzzle : pas simple.

AV : quel temps doit-on donner en communication et formation par rapport au « temps technique » ?

SC : les besoins sont aussi fonction des thématiques scientifiques (par exemple, données de santé), volumes, partage, accès,... Sur le partage, besoin de coordonner des équipes qui travaillent déjà sur le sujet. Ne pas se lancer dans des choses trop générales. Besoin de mutualiser les services et préparer le déluge de données qui arrive. Partager l'expertise, mais aussi profiter de développements faits pour d'autres disciplines. Par exemple, lien entre données et HPC, articuler différents types de traitements. Des enjeux de sobriété énergétique à l'ordre du jour, mais ce n'est pas seulement une affaire de mutualisation.

FH : D'accord avec les trois catégories citées par Odile Hologne. Les communautés Système Terre et Environnement sont diverses avec des besoins différents. Besoin d'interface avec les services EOSC. Simplifier le parcours des utilisateurs, par un accompagnement. Rôle des e-infrastructures et des data stewards. Investir dans ces compétences. Mutualiser les plateformes à l'échelle nationale. Un gros travail a déjà été accompli. Le faire à l'échelle de

l'Europe (moins évident, dans un contexte encore mal défini, mais qui n'empêche pas d'avancer). Doit être porté par les besoins des communautés. Impliquer les communautés dans la définition des services. Tester les modèles économiques et l'engagement des établissements dans la durée.

AV : ne pas tout attendre du top-down et avancer en marchant. Ecouter les communautés et faire remonter leurs besoins.

VB : Les communautés présentes dans l'EOSC sont déjà très FAIR, et les projets sont puissants et financés. La question est surtout les besoins pour la longue traine. La première étape concerne la FAIRisation des données. Informer sur les services en place, et former. Se tenir informé des services disponibles en France (établir un catalogue de services qui pourraient être connectés à EOSC) pour guider les chercheurs. Il est plus facile de partager des ressources au plan national qu'au plan européen. Etre capable de dire ce que l'Europe doit financer : compléter la connaissance partielle des experts qui travaillent sur EOSC et connaissent plutôt les communautés déjà utilisatrices et pas les autres.

AV : ne faudrait-il pas assurer un budget sur information/formation au lieu de consacrer tout l'argent au développement technique ?

Comment développer l'interdisciplinaire ?

FH : Vaste sujet aux enjeux considérables. Ce sont aux limites entre domaines que se trouveront les progrès scientifiques. Besoin de croiser les données. Les services manquent. Data Terra y travaille, par grandes communautés. Faire tomber les verrous (par exemple ZRR) qui empêchent un chercheur de consulter des données hors de sa communauté. Financement acquis pour développer des services communs à plusieurs infrastructures permettant à différents sites de discuter entre eux. Services à la demande permettant de construire un workflow sans se préoccuper du stockage des données. Gros chantiers sur l'observation spatiale. Par le passé, des couplages entre jeux de données ont amené des progrès scientifiques. Faire évoluer les pratiques des communautés, pour faire le meilleur usage des services.

Faire tomber les verrous, reconnaissance des services d'une infrastructure à l'autre.

Comment une infrastructure peut-elle gérer l'ouverture d'un service à d'autres utilisateurs.

OH : Déjà un gros travail avec des communautés bien définies (par exemple sur le blé) Quels formats de données ou de métadonnées (travail effectué dans le cadre de RDA) ? Quels changements de pratiques ? (passer de l'utilisation d'Excel sur son PC à des services mutualisés). Autre exemple : PigWeb (projet d'infrastructure de recherche sur le porc), pour partager les données sur le porc, il faut déjà harmoniser les protocoles expérimentaux, en amont du partage de données entre communautés.

De manière plus générale, du travail à faire sur les ontologies : préciser, décrire le modèle des données pour améliorer l'interopérabilité, avant d'aborder l'interdisciplinarité.

SC : Insiste sur la partie AAI et sur la transparence des accès aux différents types d'infrastructure. La formation et l'information sont des préalables. Difficile d'aller très loin sans aborder le modèle de financement. Quelle infrastructure réseau ?

AV : Intérêt d'aborder l'interdisciplinarité par le format des données : exemple du lien entre génomique et linguistique (dans les deux cas, analyse de « texte » et recherche de « mots »).

VB : Importance du travail effectué par RDA. Mieux définir le vocabulaire (en commençant par « fichier »,...). Pour ce qui est du financement, au national, EquipEx+, Recherche.Data.Gouv, ... mais les institutions, notamment grands organismes de recherche doivent comprendre qu'il n'y a pas de solution magique et qu'ils doivent investir.

2 Comment votre communauté construit-elle sa stratégie de migration vers EOSC ?

SC : construit sur l'expérience du passé (Grille de Calcul), partagé au national (France-Grilles) et à l'international (EGI, W-LCG). A permis d'entrer dans EOSC pour faire bénéficier les autres de ces développements (y compris pour la prochaine génération d'expériences, qui produiront un déluge de données. Mutualisation particules/astroparticules. Intérêt des clusters thématiques d'EOSC, en l'occurrence ESCAPE pour construire les outils de demain. L'étape suivante permettra d'intégrer ces outils dans EOSC. Le partage vers l'extérieur concerne plutôt des données très « abouties » (par opposition aux données brutes).

FH : l'offre de Data Terra est gigantesque et variée : la stratégie est différenciée selon la maturité des domaines et selon les attentes des communautés. Data-Terra est impliqué dans PILLAR et mène des projets pilotes en matière de HPC. Permet de valider une stratégie et d'explorer le coût de transférer nos services. Dans FAIREase, sélectionner des services aux interfaces, les faire murir pour les mettre à disposition d'EOSC en général et tester ce processus de migration. Pas de stratégie proprement dite, mais des essais. Les communautés ne mesurent pas encore ce que signifiera une migration vers EOSC. Besoin de se caler et de cadrer les conditions de l'ouverture par essais-erreurs.

OH : Des démarches communautaires via des projets européens, par exemple les clusters ENVRI FAIR (environnement) ou EOSC Life (génomique). INRAE est aussi embarqué dans des projets plus techniques. Par exemple dans EOSC Pillar pour i) Connecter le catalogue data.inrae au catalogue EOSC, avec le moteur de recherche B2Find ii) analyser des données via des notebooks Jupyter hébergés dans le cloud. Le concept EOSC, e-infra fédérative permettant l'interdisciplinarité, est intéressant à l'échelle de l'INRAE, d'où le travail interne pour assembler des briques et construire un « mini-EOSC à l'échelle de l'institution ».

3 Quelle relation entre EOSC et HPC ? Quelle offre de service par EURO-HPC ?

VB : différence d'approche : il n'est pas facile d'utiliser le HPC dans un contexte EOSC. C'est cependant une possibilité importante, au plan national et européen. Effort français avec FITS et MésoNet. Effort de labellisation des centres traitement données (national, régional,

mésocentres). Discussions au niveau européen sur la connexion EOSC-EuroHPC. Historiquement, les acteurs de l'HPC et ceux de l'HTC sont vraiment différents.

Philippe Segers : les choses évoluent, et le besoin de relier les deux est fort. L'arrivée de l'IA a amené à faire beaucoup évoluer les moyens d'accès (le HPC n'est plus réservé aux « gros calculs », besoin d'accès de courte durée, d'accès dynamiques).

SC : ce besoin de fluidité est important dans toutes les thématiques, et la France a du retard par rapport aux US et aux voisins européens (sur la manière d'inclure le HPC dans un flux).

4 Appels d'Offres Compétitifs de la CE : comment les institutions françaises pourront-elles y répondre ?

VB : il est nécessaire d'informer les collègues sur les appels à projets : c'est une responsabilité des institutions, qui peuvent mieux faire. Organiser des webinaires,... Nous n'avons pas une connaissance suffisante du système d'évaluation des projets au niveau des laboratoires et des universités, car nous n'avons pas assez d'experts dans les comités d'évaluation : être expert¹ permet de mieux comprendre comment écrire une proposition, c'est déterminant (quelques collègues dans chaque labo).

AV : il faut s'engager dans les GT du CoSIN, dans les TaskForces de l'EOSC-Association,... le petit nombre de français est déjà partout et fait surtout de l'entre soi. Comment aller chercher d'autres collègues ?

Philippe Segers : intérêt d'être expert pour apprendre. Par ailleurs des sociétés de service sont spécialisées dans le montage de projets européens : ils savent ce qu'il faut faire et ont une expertise que les chercheurs ne peuvent pas avoir à ce niveau.

OH : Souvent, la commission envisage de financer un tout petit nombre de projets pour chaque appel : difficile d'avoir une place, qui sont déjà préemptées par des collègues qui ont déjà une longue expérience de tels projets. Espérer au mieux un use case ou un workpackage.

VB : par exemple, le CNRS est assez grand et a l'expérience pour porter une réponse à un appel européen, de même que l'INRAé et autres (mais sur le volet HORIZON-INFRA-2022-EOSC, le jeu est fermé).

FH : pour gagner un projet, il faut s'y prendre très en avance, en équipe. Même avec l'appui d'un grand organisme et le conseil, il n'est pas facile de trouver un chef de projet : c'est un travail très prenant : pose une question de valorisation. La communauté d'experts en FAIRisation des données, en Infrastructures est assez réduite.

SC : il faut construire des communautés en amont pour répondre à ces appels. La période pour répondre est toujours plus courte. Il faut rendre des comptes et les organismes ne sont pas prêts à fournir le soutien administratif au niveau nécessaire.

¹ <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/work-as-an-expert>

FH : il faudrait mieux exploiter les caractéristiques des règles d'Horizon Europe : les établissements introduisent une rigidité qui n'existe pas dans d'autres pays (exemple d'un appui sur un partenaire privé).

AV : quel rôle pour le MESRI pour orchestrer les réponses aux appels européens ? pour « mettre en musique » ?

FH : sur EOSC, il faut miser pour voir. Au vu des constats des premiers appels, une aide au montage de futurs projets pourrait être déterminante pour que des institutions françaises portent de futurs projets.

Question sur l'interdisciplinarité :

Importance du fait que les chercheurs se parlent. Prend du temps. Il est vraiment difficile de découvrir les données d'un autre domaine scientifique quand on ne maîtrise pas le vocabulaire. C'est un énorme défi.

Baptiste Cecconi : les partenaires d'un projet européen sur « Extreme Data Mining » ignoraient l'existence d'EOSC : donne à réfléchir.