EOSC: attentes et perspectives du point de vue d'Inria

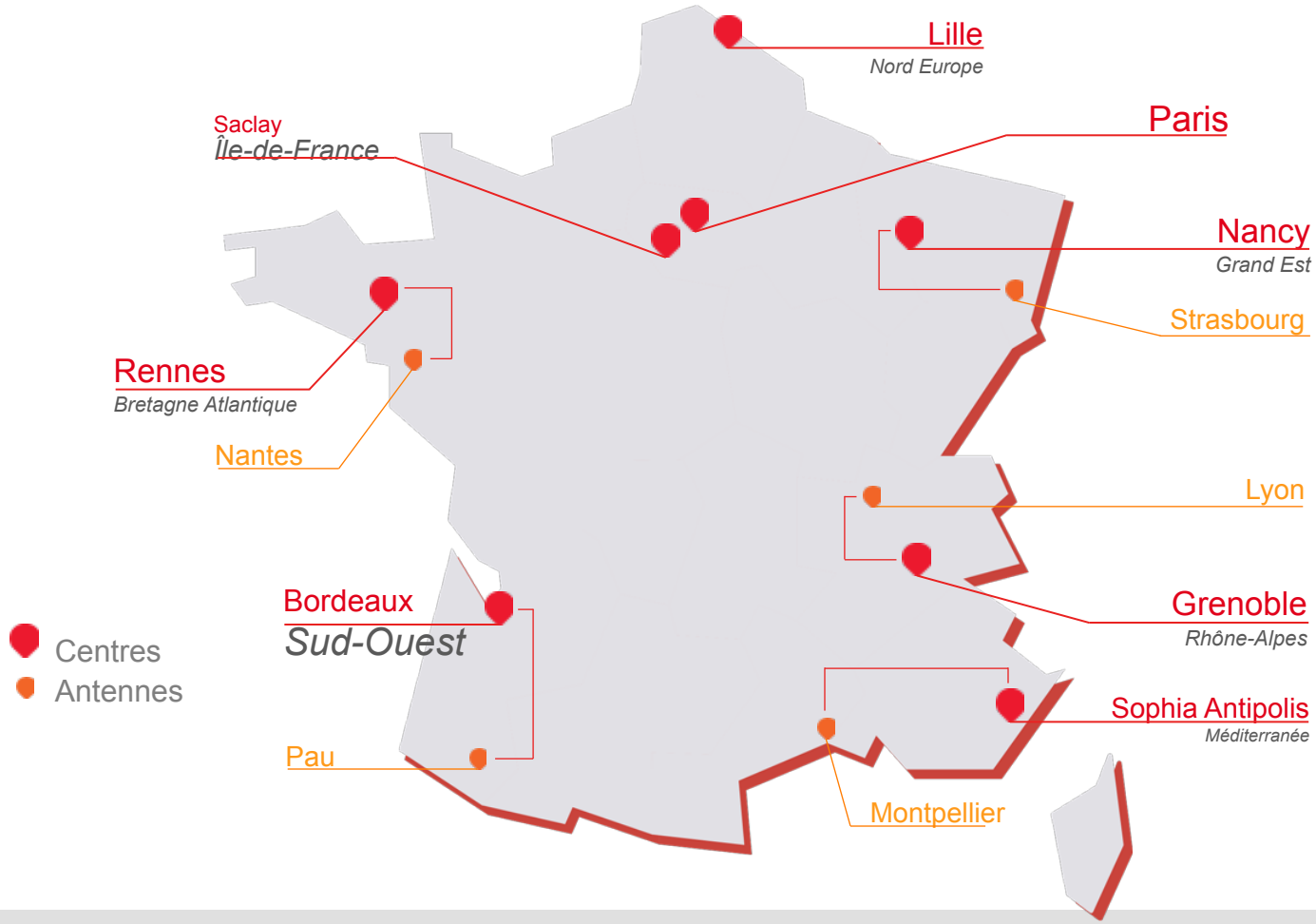
Laurent Romary, conseiller à l'IES, DGDS Inria

Directeur de recherche, équipe ALMAnaCH, Inria Paris

Différents points de vue concernant EOSC

- Une construction politique voulue par la commission européenne
  - Définir une alternative publiques aux offres privées d'hébergement et de calcul dans les nuages
- La promesse d'un réseau des **données de la recherche**
  - Hébergement, accès, calcul, services
  - Coordination de capacités nationales des états membres
- Lignes de fuite
  - Diversité des communautés représentées, e.g. projet SSHOC
  - Complémentaire des efforts menés dans le domaine des publications scientifiques
  - Une feuille de route technique qui reste à construire
- Positionnement d'Inria dans ce domaine
  - Thématiques, besoins, actions en cours

Inria – un institut, 8 implantations régionales



Informatique et mathématiques appliquées

ALGORITHMES & PROGRAMMATION

MODÉLISATION & SIMULATION

SCIENCE DES DONNÉES & INGÉNIERIE DE LA CONNAISSANCE

OPTIMISATION & CONTRÔLE

ARCHITECTURES, SYSTÈMES & RÉSEAUX

INTERACTION & MULTIMÉDIA

SÉCURITÉ & CONFIDENTIALITÉ

INTELLIGENCE ARTIFICIELLE & SYSTÈMES AUTONOMES

A l'interface d'un large spectre de thèmes socio-économiques

SANTÉ

ÉNERGIE

SÉCURITÉ &RÉSILIENCE

ENVIRONNEMENT

CLIMAT

TRANSPORT

CULTURE & DIVERTISSEMENT

ÉCONOMIE

FINANCE

ALIMENTATION & AGRICULTURE

Rappel – obligation de dépôt dans HAL

- Toute publication d'un chercheur d'une équipe Inria doit être déposée dans HAL
  - Journaux, conférences chapitres de livres
    - Même si l'article a été publié dans un journal dit « en open access » (avec paiement d'APC)
  - Lien avec les dépôts dans des archives tierces: e.g. arXiv, PMC
  - Encouragement au dépôt de preprints
- Objectif
  - Disposer d'un corpus souverain de nos publications
    - Données fiables (cf. référentiels auteurs, structures dans AureHAL)
  - En faciliter la diffusion et donc l'usage, puis la citation
  - Se conformer aux obligations de l'EU, et de l'ANR (référentiels d'AureHAL)

Enquête sur les données de la recherche

- Enquête interne Inria – ouverte à tous les personnels, toutes institutions comprises
  - Mai-juillet 2019
- Bonne participation, très bonne couverture
  - 122 réponses individuelles, 17 au titre de l'équipe
  - 115 équipes représentées, très bon équilibre entre les centres
- Bilan général
  - Variété des domaines d'application, ainsi que des formes et tailles des jeux de données
  - Conscience des enjeux : accompagnement des publications, reproductibilité, réutilisabilité
  - Fragmentation des modes de gestion et d'hébergement
  - Attente dans les domaines des PGD, de l'hébergement, du conseil juridique

Reflète la richesse des thématiques Inria

Algèbre linéaire numérique, Algorithmique arithmétique, Analyse de données, Analyse de données à grande échelle, Analyse de signaux EEG, MEG, Interfaces Cerveau Ordinateur, Analyse numérique, calcul scientifique, Modélisation, Simulation HPC, architecture des calculateurs, Assistance à la personne, Automatique, Bioinformatique, biomathématiques, microbiologie, Biologie computationnelle, biologie des systèmes, Biologie numérique, Biomécanique, Biomedical Engineering, Biostatistique, Calcul formel, Calcul intensif / HPC, Calcul parallèle, calcul scientifique HPC, Cancérologie, Compilation, Computational geometry and algebra, computer vision, Cybersécurité, data, mining, decentralised, communication networks, Distributed systems, Environnement, Évaluation et optimisation de performances de grandes infrastructures de calculs, fouille de données biomédicales, apprentissage, représentations des connaissances, Génie Logiciel, bases de données, Conception de langage, géométrie et topologie algorithmiques, Géométrie, Algèbre, Modélisation, Gestion et protection des données personnelles, Haptics, parallelisme, algorithmique, IHM , IHM visualisation, imagerie médicale, Intelligence Artificielle, optimisation, apprentissage, Intelligence artificielle, science des données, langages de programmation, les interfaces cerveau-ordinateur, Logique mathématique, Linguistique computationnelle, Machine Learning, mathématiques, mathématiques appliquées (simulation,), mathématiques appliquées et informatique, Mathématiques appliquées et simulation, électrophysiologie cardiaque, environnement, Mathématiques discrètes et codage, mathématiques, physique statistique, théorie des probabilités, statistiques, géométrie, anatomie computationnelle

Maths appliquées, Analyse et modélisation numérique, Mécanique des fluides, Propagation d'ondes,, Environnement, Calcul intensif et parallèle, Système d'information intégré, Micro-architecture, modélisation neurosciences, Modélisation probabiliste, Modélisation stochastique, modélisation stochastique, apprentissage statistique/ profond, traitement d'image, teledetection,, imagerie de la peau, modelisation/optimisation, networks, Neuroinformatique, Neurosciences Computationnelles, Optimisation, Optimisation et complémentarité, ordonnancement pour le calcul parallèle, Perception interaction cognition, Preuves et vérification, Preuves formelles, Probabilités, Problèmes inverses, production de la parole, Réalité Virtuelle, Représentation des connaissances et raisonnement automatique, Réseaux, Réseaux informatiques, sécurité, Réseaux mobiles, Robotique , Robotique et intelligence artificielle (2), Santé, biologie et planète numériques / Sciences de la planète, de l'environnement et de l'énergie, Simulation numérique, Software Engineering / Programming Languages, Statistique, Neurosciences, synthèse d'images et acquisition numérique, systems, software engineering, TAL, théorie algorithmique des nombres, Théorie de jeux appliquée aux réseaux, Theory of control, Traitement automatique des langues, Traitement automatique du langage, traitement de la parole, traitement du signal, Traitement du signal audio, Traitement du signal et apprentissage, traitement du signal et machine learning, Véhicule autonome, Véhicule autonome et robotique mobile, Vérification et Preuves Formelles, Vérification formelle de protocoles cryptographiques, Vision artificielle, apprentissage automatique, Vision par ordinateur

Origine des données

Quelle est l'origine des données ?		
Vous êtes le seul créateur des jeux de données	75	53,96%
Les données ont été créées en collaboration	88	63,31%
Les données sont issues d'autres organisations, bases de données internationales, organisations patrimoniales, données industrielles	80	57,55%
Autre	3	2,16%

Pas de profil particulier… fort taux de réponses multiples

Hébergement

Comment hébergez-vous vos données ?		
Stockage de masse local (disque dur, CD etc.) (merci de préciser)	111	79,86%
Dans une plate-forme de partage: Git, github, gitlab (merci de préciser)	75	53,96%
Dans un cloud externe (merci de préciser)	13	9,35%
Dans une archive générique telle que Zenodo (merci de préciser)	12	8,63%
En accompagnement d'une publication déposée dans une archive ouverte telle que HAL (merci de préciser)	19	13,67%
En accompagnement d'une publication disponible sur le site d'un éditeur commercial (merci de préciser)	5	3,60%
Autre	9	6,47%

CD, DVDs, serveurs d'équipe
Gitlab et github sont +très+ utilisés
Présence croissante de Zenodo, référence à Huma-Num

Montée en charge à Inria

- Publications de deux notes de cadrage:
  - Note dite courte (4 pages) : GEDEI 14299 « Note sur l'ouverture des données de la recherche »
    - Cadrage de la politique Inria en matière de gestion des données
  - Note longue (20 pages) : « La gestion des données de recherche à Inria - Guide de bonnes pratiques »
    - Détails concernant le cadre politique, technique et légal
- Premières étapes:
  - Mise en place d'une cellule nationale de contact sur les données de la recherche
    - Représentants des différentes fonctions impactées par les données de la recherche (IES, archives, DSI, DPO, FSD)
    - donnees@inria.fr
  - Espace documentaire sur l'intranet: ressources, exemples de PGD
  - Implication de l'IES dans l'accompagnement à la production de PGD
  - Implication nationale et internationale
    - HAL, SH, CoSO, EOSC
- 7 recommandations

Recommandations d'Inria pour l'ouverture des données de la recherche

- **Décrire** les jeux de données qui sont collectés, générés et analysés
- **Organiser et documenter** les jeux de données.
- Définir la **méthodologie et les standards** utilisés pour rendre ses données trouvables et interopérables
- Choisir des **supports de stockage** adaptés au projet (en termes de capacité et de coûts) et s'assurer de la sécurité des données.
- Utiliser des **licences** garantissant l'attribution et limitant le moins possible la réutilisation des données correspondantes (licence CC-by), et signaler celles qui ne sont pas librement accessibles.
- Identifier les **données sensibles** (données personnelles ou confidentielles, données susceptibles de faire l'objet d'une exploitation industrielle, ou données qui touchent à la sécurité nationale).
- **Sélectionner et archiver** les données conservées à la fin du projet.

Et le logiciel?

- Intégrer le logiciel à la réflexion sur les données
  - Une production scientifique comme une autre?
    - Reflet de nos méthodes et procédés (cf. carnets de laboratoire)
  - Elément essentiel dans la validation/réutilisation des jeux de données
    - Simulation, jeux de test, reproduction de données secondaires calculées
  - Le logiciel comme donnée de recherche (calculabilité, sécurité)
- Nécessité de penser à la préservation, identification, réutilisation du logiciel
  - Intégration dans la réflexion sur la science ouverte (e.g. licences)
  - Implication d'Inria dans Software Heritage (archive mondiale de codes sources)

Perspectives pour l'EOSC du point de vue d'Inria

- Construire autour de l'expérience acquise au niveau national
  - Retour d'expérience à partir d'infrastructures nationales telles que HAL
    - Mise en commun technique, réseau de professionnels accompagnant le déploiement
    - Interopérabilité liée à l'existence de référentiels publics communs (IdREF, AureHAL)
  - Gérer en priorité la fragmentation des solutions d'hébergement de données de la recherche
    - Cf. réflexions autour des données dites de longue traine dans le cadre du COSO
    - Anticiper un couplage étroit avec les publications – rôle de HAL dans l'écosystème des données
  - Bien intégrer la dimension du logiciel
    - Cf. Software Heritage - *Task Force on Scholarly Infrastructures for Research Software* (SIRS)
- Ne pas voir EOSC comme un enjeu purement politique
  - Défendre un projet scientifique et infrastructurel avant tout
  - Centrer la réflexion sur les communautés scientifiques (sortir du top-down FAIR)
  - Un point de vigilance: résoudre la fragmentation conceptuelle et technique pouvant être induite par la multiplicité des projets financés dans le cadre d'EOSC