Interopérabilité et pérennisation des données de la recherche

Comment FAIR En pratique ?

Livret de synthèse

Journée thématique du 27 novembre 2018
Groupe de travail inter-réseaux Atelier Données
Amphithéâtre Charpak - Campus Jussieu-Paris
https://gt-donnees2018.sciencesconf.org/

novembre 2019
Mission pour les Initiatives Transverses et Interdisciplinaires
http://www.cnrs.fr/mi/spip.php?article1313

Avertissement

Compte tenu du succès de la journée *Interopérabilité et pérennisation des données de la recherche : comment FAIR En pratique ?* organisée le 27 novembre 2018, nous n'avons pas pu retenir l'ensemble des propositions de communications reçues.

Ce livret de synthèse de la journée vise à valoriser l'engagement de chacun et contient l'ensemble des propositions de communications (résumés), celles qui ont fait l'objet d'une présentation lors de la journée (Partie I, dans l'ordre de passage des sessions) ainsi que celles qui n'ont pas pu être présentées (Partie II, par ordre alphabétique). Il permet de rendre compte et de valoriser les différentes expériences de mise en œuvre des principes FAIR dans les projets de gestion de données de la recherche.

Comité d'organisation

Renatis - Emmanuelle Morlock (coordinatrice)
Médici - Caroline Martin (coordinatrice) / Stéphane Renault
QeR - Alain Rivet
Resinfo - Olivier Brand-Foissac / Maurice Libes
rBDD - Marie-Claude Quidoz / Geneviève Romier
Calcul - Anne Cadiou / Loïc Gouarin
Devlog - Pierre Brochard / Dominique Desbois
Représentant de la DIST-CNRS - Joanna Janik

Contact
gt-donnees-inter-reseaux@groupes.renater.fr

Mise en page : Stéphane Renault - LaMPEA (UMR 7269) / réseau Médici
Livret en accès libre : https://gt-donnees2018.sciencesconf.org/

Introduction

Avec le mouvement de l'ouverture des données de la recherche et du partage dit "Science Ouverte", au niveau européen et international, de nombreux pays se sont engagés dans des politiques permettant le partage des résultats scientifiques et des données. La Science Ouverte est une nouvelle approche de la démarche scientifique, basée sur la production collaborative des produits de science, de leur partage, de leur libre circulation et réutilisation. La Commission européenne a émis dès 2012 des recommandations concernant la diffusion des résultats scientifiques. Elles invitent les chercheurs à s'appuyer sur les principes FAIR (Findable, Accessible, Interoperable, Reusable) permettant de favoriser la découverte, l'accès, l'interopérabilité et la réutilisation des données. Ils permettent de guider les stratégies de gestion des données et d'aider tous les acteurs qui œuvrent à les produire, à en contrôler la qualité, à les traiter et les analyser, à assurer leur publication et leur dissémination, à les sélectionner et les préparer pour le dépôt dans des plateformes de partage ou d'archivage.

Organisée par le groupe de travail inter-réseaux « Atelier Données » de la Mission pour les Initiatives transverses et interdisciplinaires (MITI), la journée d'étude a pour objectif de présenter des retours d'expériences et des réflexions sur les pratiques de gestion des données de la recherche mises en œuvre par les réseaux métiers et les réseaux technologiques du CNRS. Elle s'appuie plus spécifiquement sur les notions de pérennisation et d'interopérabilité des données dans les projets de recherche, d'en comprendre les facteurs ressorts de réussite et les points sensibles à surveiller.

L'ambition est :

- d'analyser les complémentarités de ces expériences au travers des métiers représentés par les réseaux ;
- de formuler des points de convergence de bonnes pratiques ;
- d'accroître les échanges entre les réseaux de la MITI sur des questions à forts enjeux pour l'évolution de nos métiers.

Cette journée est organisée avec le soutien de la Mission pour les Initiatives transverses et interdisciplinaires (MITI-CNRS), de la DIST du CNRS, des réseaux Calcul, Devlog, Frantiq, Médici, rBDD, Renatis, Resinfo, QeR et du Consortium MASA.

Site de la journée 2018 : https://gt-donnees2018.sciencesconf.org/

Programme de la journée du 27 novembre 2018

Introduction de la journée

- Ouverture - **Groupe de travail inter-réseaux - Atelier Données**
- Introduction - **Catherine Clerc** (responsable de la Plateforme des Réseaux – Mission pour les Initiatives transverses et interdisciplinaires (MITI)
- Les enjeux des principes FAIR - **Volker Beckmann** (Directeur Adjoint Scientifique en charge du domaine « Calcul et données » de l'Institut national de physique nucléaire et de physique des particules - IN2P3)

Session 1 : Quelles formes de pérennisation pour quels types de données ?

(Présidente de séance : Caroline Martin - GT Atelier Données, réseau Médici)

- Traçabilité des activités de recherche dans les unités de recherche - **Alain Rivet** (Centre de Recherches sur les Macromolécules Végétales)
- Retour d'expériences sur la publication de données en biologie - **Pierre Poulain** (Institut Jacques Monod)
- Pérennisation de logiciels de la recherche, le projet PRESOFT : Preservation for REsearch SOFTware ou Pourquoi et comment FAIR un SMP ? - **Teresa Gomez-Diaz** (Laboratoire d'Informatique Gaspard-Monge) et **Geneviève Romier** (Centre de Calcul de l'IN2P3)
- Les principes FAIR appliqués aux sauvegardes sur le long terme - **Marie-Claude Quidoz** (Centre d'Écologie Fonctionnelle et Évolutive)

Session 2 : Méthodes et outils, quelle alliance pour réussir les projets de valorisation des données en SHS ?

(Président de séance : Stéphane Renault - GT Atelier Données, réseau Médici)

- FAIR en linguistique de la langue orale : objectifs, méthode et outils - **Loïc Liégeois** (Laboratoire de Linguistique Formelle), **Carole Etienne** (Interactions, Corpus, Apprentissages, Représentations) et **Christophe Parisse** (Modèles, Dynamiques, Corpus)
- Les périodiques du Muséum à l'heure du tout-numérique - **Emmanuel Côtez**, **Anne Mabille** et **Chloë Chester** (Muséum d'Histoire Naturelle)
- Produire et diffuser des métadonnées thématiques pour l'archéologie : entre savoir-FAIR et pratique collaborative - **Blandine Nouvel** (Centre Camille Jullian) et **Miled Rousset** (Maison de l'Orient et de la Méditerrannée Jean Pouilloux)
- Les principes FAIR appliqués aux données archéologiques produites par l'Inrap : état de la réflexion et des projets en cours - **Emmanuelle Bryas**, **Camille Colin**, **Anne Moreau** et **Christophe Tufféry** (Institut National de Recherches Archéologiques Préventives)

Discussion

- Échanges avec les participants

Session 3 : Les principes FAIR, une nouvelle opportunité pour améliorer ses pratiques en matière de gestion de données ?

(Président de séance : Maurice Libes - GT Atelier Données, réseau Resinfo)

- Observatoire Virtuel et FAIR, des principes fondamentaux à la pratique - **André Schaaff** (Observatoire Astronomique de Strasbourg), **Laurent Bourgès** (Observatoire des Sciences de l'Univers de Grenoble), **Karin Dassas** (Institut d'Astrophysique Spatiale), **Jean-Michel Glorian** (Institut de Recherche en Astrophysique et Planétologie), **Jean-Charles Meunier** (Laboratoire d'Astrophysique de Marseille), **Michèle Sanguillon** (Laboratoire Univers et Particules de Montpellier) et **Pierre Le Sidaner** (Observatoire de Paris)
- Les documentalistes et la FAIRisation des données scientifiques : un travail d'équipe inter-métiers - **Soizick Lesteven** (Observatoire Astronomique de Strasbourg)
- Pérennisation et interopérabilité des données de l'Observatoire de Recherche Méditerranéen de l'Environnement - **Juliette Fabre** et **Olivier Lobry** (Observatoire des Sciences de l'Univers - OREME)
- Make your data great (again) - **Daniel Jacob** (Institut National de la Recherche Agronomique - Inra)
- OpenData pour la simulation et les expérimentations au LEGI - **Gabriel Moreau**, **Antoine Mathieu**, **Cyrille Bonamy**, **Joël Sommeria** et **Julien Chauchat** (Laboratoire des Écoulements Géophysiques et Industriels)

Session 4 : Synthèse de la journée

(Présidente de séance : Emmanuelle Morlock - GT Atelier Données, réseau Renatis)

- Présentation de RDA France - **Francis André** (Direction de l'Information Scientifique et Technique du CNRS)
- Conclusion synthèse journée - **Volker Beckmann** (Directeur Adjoint Scientifique en charge du domaine « Calcul et données » de l'Institut national de physique nucléaire et de physique des particules - IN2P3)

Discussion

- Échanges avec les participants

La journée a été retransmise en direct par webcast

Les enregistrements sont disponibles sur la plateforme de webcast du CC IN2P3
https://webcast.in2p3.fr/container/interoperabilite-et-perennisation-des-donnees-de-la-recherche

Partie I - Session 1
Interventions du 27 novembre

Quelles formes de pérennisation pour quels types de données ?

- Traçabilité des activités de recherche dans les unités de recherche
  Alain Rivet (Centre de Recherches sur les Macromolécules Végétales)
- Retour d'expériences sur la publication de données en biologie
  Pierre Poulain (Institut Jacques Monod, UMR7592, CNRS et Université Paris Diderot)
- Pérennisation de logiciels de la recherche, le projet PRESOFT : Preservation for REsearch SOFTware ou Pourquoi et comment FAIR un SMP ?
  Teresa Gomez-Diaz (Laboratoire d'Informatique Gaspard-Monge)
  et **Geneviève Romier** (Centre de Calcul de l'IN2P3)
- Les principes FAIR appliqués aux sauvegardes sur le long terme
  Marie-Claude Quidoz (Centre d'Écologie Fonctionnelle et Évolutive)

Traçabilité des activités de recherche dans les unités de recherche

Alain Rivet
Centre de Recherches sur les Macromolécules Végétales, Grenoble / réseau Qualité en Recherche

La mission d'un organisme de recherche tel que le CNRS consistant à produire et valoriser des connaissances, l'information qui est alors générée s'avère un patrimoine essentiel qu'il convient de préserver. La gestion des données de la recherche qui s'inscrit dans le cadre de la « Science Ouverte » vise à rendre accessible les résultats de la recherche scientifique et s'avère essentielle pour assurer la fiabilité du travail de recherche, élément clé d'une démarche qualité en recherche.

Le réseau « Qualité en Recherche » de la Mission pour les Initiatives Transverses et Interdisciplinaires a mis en place un groupe de travail intitulé « Traçabilité des activités de recherche et gestion des connaissances dans les unités de recherche » destiné à apporter une vision « qualité » à cette thématique. Cette initiative s'articule avec le projet national « Stratégie de conservation des données scientifiques et administratives au CNRS » piloté conjointement par la MPR (Mission pilotage et relations avec les délégations régionales et les instituts) et le pôle national de conservation des données et documents de la DAJ (Direction des Affaires Juridiques).

Le guide « Traçabilité des activités de recherche et gestion des connaissances - Guide pratique de mise en place » librement accessible sur le site du réseau Qualité en Recherche[1], se propose d'être une réponse à la nécessité d'assurer la maîtrise des données de la recherche face à cette explosion des données et ce, dans le respect des contraintes administratives et règlementaires (loi pour une République numérique, RGPD…).

Le guide a comme objectif de fournir des recommandations et bonnes pratiques pouvant être appliquées dans tous les domaines d'activités, tant administratifs, techniques que scientifiques, afin d'assurer la traçabilité des activités de recherche et d'améliorer la gestion des connaissances de nos structures de recherche.

Ces recommandations, partant de l'expérience de plusieurs unités de recherche, sont essentiellement organisationnelles et peuvent s'intégrer dans la mise en place de démarches qualité au sein de structures de recherche. Également destinées à rendre les données « FAIR », ces recommandations dont l'organisation suivante est proposée, consistent à :

- disposer d'outils d'enregistrement et de traçabilité ;
- identifier efficacement les fichiers numériques ;
- définir un plan de classement des dossiers ;
- créer un plan de gestion de données (PGD) ;
- sélectionner les données ;
- sauvegarder et archiver les données ;
- communiquer et sensibiliser le personnel.

[1] http://qualite-en-recherche.cnrs.fr/IMG/pdf/guide_tracabilite_activites_recherche_gestion_connaissances.pdf

Retour d'expériences sur la publication de données en biologie

Pierre Poulain
Institut Jacques Monod (UMR7592, CNRS et Université Paris Diderot), Paris

Il s'agit de dresser un rapide inventaires des moyens disponibles pour publier des données issues de la recherche en biologie.

Publication d'un tableau de données comme données complémentaires à un article scientifique (*supplementary data*)

Ex. : Poulain *et al.*, PLOS One, 2010 (DOI 10.1371/journal.pone.0009990) avec les données[1]. Cette méthode est sans doute la plus simple mais ne permet pas toujours de réutiliser facilement les données ainsi publiées, notamment si le format de données n'est pas pertinent (PDF par exemple). Un *Digital Object Identifier* (DOI) n'est pas systématiquement affecté à ce type de données. Il faut également être vigilant à la licence d'utilisation de ces données.

Publication de paramètres de simulation via un dépôt de données externe : le cas Figshare

Ex. : Jallu *et al.*, PLOS One, 2012, (DOI 10.1371/journal.pone.0047304) avec les données[2]. Les données sont stockées sur un site internet externe (c'est-à-dire qui n'est pas le site du journal scientifique ni celui d'un des auteurs). Un DOI est attribué aux données. Figshare semble respecter les principes FAIR[3]. Mais que penser des liens entre Figshare et l'éditeur Springer Nature *via* le groupe Holtzbrinck[4] ? La licence Creative Commons Attribution (CC-BY) largement utilisée pour la publication des données dans Figshare suffit-elle à garantir la pérennité des données ?

Publication d'un tableau données via un dépôt de données externe : le cas Zenodo

Ex. : Etoka-Beka *et al.*, Tropical Medicine & International Health, 2016 (DOI 10.1111/tmi.12786) avec les données[5]. Zenodo est hébergé par l'organisation européenne pour la recherche nucléaire (CERN) et respecte complètement les principes FAIR[6]. Les données sont associées à un DOI et une licence d'utilisation.

Publication de données de la recherche un peu particulières : le code informatique

Ex. : Barnoud *et al.*, PeerJ, 2018 (DOI 10.7717/peerj.4013) avec les données[7].

Le code informatique est déposé sur la plateforme de développement GitHub mais également automatiquement archivé sur Zenodo. Le choix de la licence associée au code informatique (toutes ne se valent pas) est important. Que penser de GitHub racheté par Microsoft en juin 2018 ? Zenodo apparaît alors comme un bon complément. L'initiative *Software Heritage*, dont l'objectif est de préserver et archiver tout le code informatique disponible publiquement, est aussi particulièrement intéressante.

Publication de données biologiques « massives »

Aujourd'hui, la publication de données biologiques « massives », issues notamment de la génomique et de la protéomique, est devenu un standard. Des sites comme *Gene Expression Omnibus* (GEO)[8], *Sequence Read Archive* (SRA)[9] ou *PRoteomics IDEntifications* (PRIDE)[10] sont des références dans ces domaines. L'utilisation de ces dépôts est recommandée voir imposée par les éditeurs scientifiques. Néanmoins, la conformité de ces dépôts par rapport aux principes FAIR n'est pas toujours évidente.

---

1 https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009990#s5

2 https://figshare.com/articles/Model_of_the_beta3_subunit_of_integrin_alphaIIb_beta3/104602 et https://figshare.com/articles/Molecular_Dynamics_Protocol_with_Gromacs_4_0_7/104603

3 https://figshare.com/articles/Figshare_and_FAIR_data_principles/6505655

4 https://scholarlykitchen.sspnet.org/2017/10/23/ownership-digital-science/

5 https://zenodo.org/record/154453#.W6DfOXUzZhE

6 http://about.zenodo.org/principles/

7 https://github.com/pierrepo/PBxplore et https://zenodo.org/record/1016257

8 https://www.ncbi.nlm.nih.gov/geo/

9 https://www.ncbi.nlm.nih.gov/sra

10 https://www.ebi.ac.uk/pride/archive/

Pérennisation de logiciels de la recherche Le projet PRESOFT : Preservation for REsearch SOFTware ou Pourquoi et comment FAIR un SMP ?

Teresa Gomez-Diaz
Laboratoire d'Informatique Gaspard-Monge-LIGM, Champs-sur-Marne

Geneviève Romier
Institut national de physique nucléaire et de physique des particules-IN2P3, Villeurbanne

Réutilisation, Reproductibilité de la science, Pérennisation de logiciels de la recherche

PRESOFT : Preservation for REsearch SOFTware est un projet d'étude et de prototypage des plans de développement logiciels pour les logiciels de la recherche développés dans les laboratoires.

Alors que les plans de gestion de données sont maintenant exigés par les agences de financement dans la plupart des projets, il n'est pas encore très fréquent qu'un plan de gestion du logiciel (Software Management Plan - SMP) soit demandé explicitement dans le cadre d'un appel à projet. Pourtant, les logiciels de la recherche jouent un rôle fondamental puisque la plupart des résultats scientifiques s'appuient sur des analyses de données, simulations, ou calculs obtenus grâce à ces logiciels. Par ailleurs, un SMP est, de même qu'un plan de gestion de données, un bon outil pour mieux gérer la production scientifique sur le moyen et le long terme.

PRESOFT est un projet (2017-2018) CNRS entre deux laboratoires IN2P3 et le LIGM dont les objectifs sont :

- de développer des procédures et modèles réalistes pour les plans de gestion du logiciel qui pourraient être proposés dans les laboratoires ;
- d'étudier la faisabilité, les bénéfices par rapport aux contraintes, l'acceptabilité et les conditions pour une réelle adoption par les chercheurs, les ingénieurs, les thésards ou les responsables des projets de développement logiciel pour leur propre production de logiciels de recherche de plans de gestion du logiciel ;
- d'évaluer l'impact pour une unité de l'implémentation de plans de gestion du logiciel sur sa connaissance du logiciel de recherche développé en son sein et sa gestion. Par exemple, sur la gestion des actifs logiciels internes, le coût de la pérennisation du logiciel, sa valorisation, les compétences internes...

Un modèle de plan de gestion du logiciel finalisé avec l'aide d'utilisateurs a été publié (http://www.france-grilles.fr/presoft/) et est intégré dans la plate-forme DMP OPIDoR de l'INIST. Plusieurs plans de gestion du logiciel ont été établis pour différents types de logiciels de la recherche. Nous pourrons présenter le contexte international et le projet, expliquer le modèle et son utilisation et développer l'intérêt pour un projet logiciel de réaliser un tel plan. Enfin, les premiers retours d'expérience et résultats de l'étude seront exposés.

Les principes FAIR appliqués aux sauvegardes sur le long terme

Marie-Claude QUIDOZ
Centre d'Écologie Fonctionnelle et Évolutive, Montpellier / réseau Base De Données

Pérennisation des données, ontologies de domaine, système de gestion de bases de données, formats de fichiers, sens des données, obsolescence

Les données qui ne semblent pas avoir vocation à être associées à des publications au moment de leur acquisition ou de leur archivage, peuvent se révéler à long terme d'un grand intérêt pour la communauté qui les a produites ou pour d'autres communautés. Mettre en place des mécanismes pour en assurer leur réutilisation est indispensable.

Cela pose un certain nombre de questions :

- Comment assurer la sauvegarde des données à long terme pour permettre leur réutilisation ?
- Les solutions proposées par les plateformes de dépôt ou d'archivage pérenne sont-elles à implémenter dans ce cas ?

La présentation tentera d'apporter des réponses à ses questions en démontrant les apports d'un système d'information organisé autour d'un système de gestion de bases de données et d'une ontologie de domaine.

Après une illustration des 4 risques menaçant un fichier sur une longue période (obsolescence matérielle, obsolescence logicielle, obsolescence du format de fichiers, perte de signification du contenu), nous nous attacherons à proposer des solutions pratiques et accessibles à mettre en œuvre du point de vue technique. Une attention particulière sera mise sur les moyens humains (veille technologique / documentation / migration) et financiers (personnel / matériel /logiciel) comme condition nécessaire à l'application de ces solutions. D'autre part, nous nous attarderons sur les formats de fichier en s'appuyant sur les travaux du CINES (outil Facile), de la TGIR Huma-Num et sur notre propre expérience (rBDD).

En définitive, tous ces outils ne prendront leur sens que si ils s'inscrivent dans une démarche FAIR adossée à un plan de gestion de données.

La présentation se conclura sur un retour d'expérience relatant l'apport des ontologies de domaine pour réduire la perte de sens des données en biodiversité.

La présentation illustrera les principes FAIR mis en œuvre au sein de la communauté : grâce au système d'information, aux plans de gestion de données, à l'utilisation d'ontologies, à l'actualisation dans le temps des formats de fichiers et protocoles, les équipes de recherche pourront trouver les données, y accéder, les réutiliser et les données seront interopérables.

Partie I - Session 2
Interventions du 27 novembre

Méthodes et outils, quelle alliance pour réussir les projets de valorisation des données en SHS ?

- FAIR en linguistique de la langue orale : objectifs, méthode et outils
  Loïc Liégeois (Laboratoire de Linguistique Formelle), **Carole Etienne** (Interactions, Corpus, Apprentissages, Représentations) et **Christophe Parisse** (Modèles, Dynamiques, Corpus)
- Les périodiques du Muséum à l'heure du tout-numérique
  Emmanuel Côtez, **Anne Mabille** et **Chloë Chester** (Muséum d'Histoire Naturelle)
- Produire et diffuser des métadonnées thématiques pour l'archéologie : entre savoir-FAIR et pratique collaborative
  Blandine Nouvel (Centre Camille Jullian) et **Miled Rousset** (Maison de l'Orient et de la Méditerrannée Jean Pouilloux)
- Les principes FAIR appliqués aux données archéologiques produites par l'Inrap : état de la réflexion et des projets en cours
  Emmanuelle Bryas, **Camille Colin**, **Anne Moreau** et **Christophe Tufféry** (Institut National de Recherches Archéologiques Préventives)

Discussion
- Échanges avec les participants[1]

[1] Les échanges avec les participants n'ayant pas fait l'objet d'une retranscription, vous pouvez les retrouver sur la captation *Les principes FAIR appliqués aux données archéologiques produites par l'Inrap : état de la réflexion et des projets en cours* (E. Bryas *et al.* - intervention de Christophe Tufféry) à partir de : 09'05"

FAIR en linguistique de la langue orale. Objectifs, méthode et outils

Loïc Liégeois
CLILLAC-ARP et LLF, Université Paris Diderot

Carole Etienne
CNRS, ICAR, ENS de Lyon

Christophe Parisse
Inserm, Modyco, Université Paris Nanterre

Si de plus en plus de corpus de linguistique de la langue orale ont été rendus disponibles dans des plateformes ouvertes ces dernières années (ORTOLANG, COCOON…) et permettent de ce fait un partage et une réutilisabilité des données de la recherche, il n'en demeure pas moins que cette réutilisation par des chercheurs d'autres équipes de recherche, voire d'autres communautés scientifiques, pose encore problème. D'une part, la sélection des données et leur documentation au moment des analyses requièrent des métadonnées qui répondent à des problématiques de recherche et pas seulement d'archivage. D'autre part, la structuration et le format des données issues d'origines diverses devraient permettre une exploitation par les différents logiciels pour différentes finalités scientifiques, sans recourir à des développements informatiques parcellaires, spécifiques et par conséquent difficiles à partager et à maintenir.

Pour répondre à ces problèmes, le groupe de travail InterExplo du consortium CORLI, piloté par Huma-Num, a proposé une solution pour faciliter et encourager la mutualisation, l'échange et l'analyse des corpus oraux et multimodaux.

Dans un premier temps, le groupe de travail a défini un jeu commun de métadonnées, répondant aux besoins des chercheurs travaillant sur l'oral et pensé comme le niveau minimum requis pour permettre une réutilisation de qualité des corpus oraux à des fins de recherche. Ensuite, il a choisi de représenter ce jeu de données dans un standard largement utilisé en linguistique et en Europe, la TEI (Text Encoding Initiative, TEI Consortium, 2018), et a proposé son modèle dans le format ODD. Enfin, une interface paramétrable a été développée pour permettre de saisir les métadonnées à partir de ce modèle et de les exporter dans un format TEI commun et partagé par la communauté.

Dans un second temps, il a conçu et développé un second outil, nommé Conversion (Parisse et al., 2017, cf. http://ct3.ortolang.fr/tei-corpo/), pour convertir les principaux formats de corpus sans perte d'informations, facilitant ainsi les analyses multi-niveaux d'un même corpus et la collaboration entre les équipes de recherche. De la même manière que pour les métadonnées, un export en TEI (TEI-CORPO) permet au linguiste de disposer d'une version de son corpus structurée dans un format ouvert, documenté, pérenne et interopérable.

Notre intervention se focalisera sur l'approche méthodologique qui a été suivie au sein du consortium CORLI, pour rassembler et prendre en compte les besoins de la communauté en amont, puis pour diffuser ces réalisations et informer les différentes équipes. Nous aborderons également les échanges avec la communauté TEI et l'évolution de ce standard pour l'oral à laquelle nous avons contribué.[1]

[1] Quelques références :

Liégeois L., Etienne C., Parisse C., Benzitoun C. et Chanard C. (in press), Using the TEI as pivot format for oral and multimodal language corpora, *Journal of the Text Encoding Initiative*, 10, halshs-01357343.

Parisse C., Benzitoun C., Etienne C., Liégeois L., 2017, *Agrégation automatisée de corpus de français parlé*, 9e Journées Internationales de la Linguistique de Corpus, Grenoble, halshs-01636957

Liégeois L., Parisse C., Etienne C., 2017, *Atelier Métadonnées*, 9e Journées Internationales de la Linguistique de Corpus, Grenoble

Liégeois L., Etienne C., Benzitoun C. & Parisse C., 2017, Vers un format pivot commun pour la mutualisation, l'échange et l'analyse des corpus oraux, *FLORAL : Accessibilité, représentations et analyses des données*, Orléans, halshs-01636964

Liégeois L., Etienne C., Parisse C., Benzitoun C. et Chanard C., 2016 , Utilisation d'un format commun pour structurer les métadonnées de corpus oraux : objectifs, enjeux et méthode, *Données, métadonnées des corpus et catalogage des objets en sciences humaines et sociales*, Poitiers, halshs-01357271

The TEI Consortium. (2018). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.3.0. L. Burnard & S. Bauman, dir. Charlottesville: Text Encoding Initiative Consortium.

Les périodiques du Muséum à l'heure du tout numérique

Emmanuel Côtez, Anne Mabille, Chloë Chester
Pôle périodiques du Muséum National Histoire Naturelle, Paris

Le Muséum est, depuis 1802, éditeur scientifique dans le domaine des Sciences de la Nature et de l'Homme. Il publie des résultats scientifiques originaux dans ces domaines de compétence, allant de l'Anthropologie à la Zoologie en passant par la Botanique, l'Ethnologie et la Paléontologie.

Sur les six revues publiées par le Service des Publications scientifiques, quatre sont consacrées à la taxonomie, cœur historique de mission de conservation de la biodiversité de l'établissement. La taxonomie est l'étude des taxons animaux et végétaux et a pour objectif de documenter, décrire et synthétiser tous les aspects de la biodiversité.

La taxonomie n'est pas une science qui oublie, et la durée de vie des articles et monographies va bien au-delà des durées habituelles dans d'autres champs de recherche. Ainsi, les publications de Linné, en1753, marquent le début d'une règle encore en vigueur aujourd'hui : sans publication, pas de description valide de nouvelle espèce, pas d'opération nomenclaturale reconnue par la communauté de taxonomistes...

Les sources sont donc éparpillées, dans le temps mais aussi dans la littérature. Pour rendre accessible celle-ci, la communauté s'est attachée dans un premier temps à numériser massivement les articles publiés depuis 1753. L'organisation des données ainsi récoltées est un chantier qui commande des procédures spécifiques.

Pour répondre à ces impératifs, les périodiques du Muséum sont passés, depuis le 1er janvier 2018, à une publication en ligne uniquement, mettant à la disposition de la communauté scientifique des articles en Open Access diamant sous forme de fichiers enrichis.

En même temps, l'équipe se tourne vers l'avenir en consacrant des efforts sur les questions de l'interopérabilité et la pérennisation des articles, ainsi que les données qu'ils contiennent – le dernier défi dans cette époque de big data où les technologies évoluent et se transforment à grande vitesse.

Emmanuel Côtez, Anne Mabille et Chloë Chester – de l'équipe du pôle périodiques du Muséum – expliqueront quelles procédures ont été mises en place pour enrichir les fichiers de publication, et avec quelle information ? Comment répondre aux impératifs de pérennisation des articles qui doivent être accessibles encore dans 200 ans ? Ils aborderont aussi la problématique de la mise en sens de ces contenus. Ainsi, la transformation en XML des articles permettra d'améliorer encore leur diffusion, leur pérennisation, et leur capacité à être exploités par la communauté scientifique. En inventant de nouvelles manières de travailler avec les données de la recherche, ce sont de nouveaux résultats qui seront découverts, puis publiés, ajoutant ainsi de nouvelles briques à la grande construction de la chaîne du vivant.

Ils exposeront enfin les enjeux de la mise en place d'une stratégie internationale intégrant une réflexion collective de tous les acteurs de la recherche, aussi bien en botanique qu'en paléontologie ou en zoologie, disciplines qui répondent à des règles et des usages parfois très divergents. Mais qui imposent aussi une réflexion de plusieurs corps de métiers : informaticiens, taxonomistes, éditeurs, bibliothécaires…

Produire et diffuser des métadonnées thématiques pour l'archéologie. Entre savoir-FAIR et pratique collaborative

Blandine Nouvel
Centre Camille Jullian-CCJ (AMU, CNRS, Minist Culture) / Fédération et ressources sur l'Antiquité-FRANTIQ (GDS 3378), Aix-en-Provence

Miled Rousset
Maison de l'Orient et de la Méditerranée-MOM (CNRS, Univ Lyon 2), Lyon

Réseau DEVLOG

Construit au départ pour les besoins du Catalogue Collectif Indexé de Frantiq, le thésaurus PACTOLS pour l'archéologie a évolué au fur et à mesure de son développement pour se trouver en capacité de répondre à d'autres besoins que ceux strictement bibliothéconomiques.

Multilingue et respectueux des standards (ISO 25964, XML-SKOS) PACTOLS est citable jusqu'au moindre concept par l'attribution d'identifiants pérennes (Ark). Il est géré avec le logiciel libre de gestion de thésaurus Opentheso, ce qui lui confère de fait des fonctionnalités propres à satisfaire les impératifs d'interopérabilité et d'accessibilité des principes FAIR du web sémantique. Hors la mue technologique aujourd'hui finalisée, nous avons initié l'année dernière une ambitieuse opération de consolidation de la structure du thésaurus PACTOLS qui s'opère sur plusieurs points : une réorganisation des domaines pour donner au thésaurus une meilleure lecture de son contenu lexical et faciliter son utilisation par une communauté d'usagers élargie, un enrichissement terminologique par l'ajout de vocabulaires de spécialités pour l'adapter au plus près des besoins des scientifiques, des alignements avec des vocabulaires de référence pour une inscription efficace dans le web de données.

L'objectif est d'offrir à la communauté scientifique un vocabulaire contrôlé spécialisé, évolutif, interfaçé avec plusieurs systèmes de gestion de contenus pour l'ajout dynamique de métadonnées thématiques (chaîne Métopes pour la production éditoriale en XML-TEI, bibliothèque numérique Omeka par exemple), et déjà inscrit dans le web sémantique, qui constitue à terme un pivot sémantique pour l'archéologie française. Ainsi, plutôt que de développer de son côté des listes de termes, le chercheur bénéficie d'un outil validé par la communauté et immédiatement opérationnel.

Nous évoquerons les difficultés rencontrées lors de ces changements pour l'équipe en charge de l'administration du thésaurus et du développement du logiciel (méthodologie, organisation du travail, moyens). Nous présenterons aussi, à travers quelques expériences d'utilisation du thésaurus, celles auxquelles se sont confrontés les nouveaux usagers, principalement des éditeurs et des chercheurs : le travail de qualification des données devient collaboratif et dépendant d'outils partagés dont ils n'ont pas une complète maîtrise. Au final, ce sont certainement les actions de formation et d'écoute réciproque qui permettront de passer d'un stade expérimental à la construction d'un réseau d'utilisateurs-producteurs où données et métadonnées seront idéalement valorisées.

Les principes FAIR appliqués aux données archéologiques produites par l'Inrap. État de la réflexion et des projets en cours

Emmanuelle Bryas, Camille Colin, Anne Moreau, Christophe Tufféry
Diection Scientifique et Technique
Institut national de recherches archéologiques préventives-Inrap, Paris

Jusqu'à une date récente, les pratiques numériques des archéologues de l'Inrap se sont appuyées sur des outils très variés et hétérogènes, conçus sans souci d'interopérabilité technique ni sémantique.

Conscient de devoir faire évoluer ces pratiques, l'Inrap a récemment engagé plusieurs actions en vue de rendre interopérables les données produites depuis le terrain jusqu'à leur publication, en s'appuyant notamment sur l'utilisation de métadonnées et de formats d'échanges normés. Cela devrait favoriser la production de nouvelles connaissances par l'analyse croisée des données archéologiques, permettre le renouvellement des modes de publication et faciliter les échanges avec ses partenaires scientifiques et institutionnels.

Ainsi, le catalogue en ligne Dolia, permet la signalisation dans le respect du format international d'échanges bibliographique Unimarc, la description à l'aide d'une indexation en langage contrôlé avec les thésaurus Pactols du réseau Frantiq et la localisation des publications acquises mais aussi des documents produits au cours des opérations archéologiques. Le catalogue Dolia permet en outre au personnel de l'Institut d'accéder en texte intégral aux rapports d'opérations, fouilles et diagnostics.

Ces mêmes rapports d'opération, dotés de métadonnées Dublin Core, font progressivement l'objet d'un archivage pérenne au CINES.

En amont, pour l'acquisition des données de terrain, l'Inrap développe depuis 2015, une application baptisée EDArc. Cette application s'appuie sur un noyau de données minimales nécessaires à enregistrer pour tous les types d'opérations d'archéologie préventive. EDArc permet de renseigner des métadonnées issues de la norme ISO 15836 et utilise certains des *thesauri* de Pactols et des listes de vocabulaires contrôlés de l'application Patriarche du Ministère de la Culture. EDArc a fait l'objet d'un appariement avec la norme ISO 21127 du CIDOC CRM, qui augure de la possibilité de publier les données de terrain dans le Web sémantique, dès lors que des décisions seront prises sur les modalités de la publication des données d'archéologie préventive en données ouvertes.

L'Inrap a mis en œuvre un catalogue de données spatiales ayant pour objectif de mettre à disposition des archéologues de l'Inrap des données spatiales, neutres et interopérables, susceptibles d'être réutilisées. Constitué peu à peu par l'accumulation des données spatiales collectées sur le terrain de manière systématique et structurée, le catalogue vise à devenir l'une des portes d'entrée vers les rapports et les lots d'archives. Dans cette perspective, un travail de définition des métadonnées a été initié, conformément aux principes de la directive INSPIRE.

Partie I - Session 3
Interventions du 27 novembre

Les principes FAIR, une nouvelle opportunité pour améliorer ses pratiques en matière de gestion de données ?

- Observatoire Virtuel et FAIR, des principes fondamentaux à la pratique
  André Schaaff (Observatoire Astronomique de Strasbourg), **Laurent Bourgès** (Observatoire des Sciences de l'Univers de Grenoble), **Karin Dassas** (Institut d'Astrophysique Spatiale), **Jean-Michel Glorian** (Institut de Recherche en Astrophysique et Planétologie), **Jean-Charles Meunier** (Laboratoire d'Astrophysique de Marseille), **Michèle Sanguillon** (Laboratoire Univers et Particules de Montpellier) et **Pierre Le Sidaner** (Observatoire de Paris)
- Les documentalistes et la FAIRisation des données scientifiques : un travail d'équipe inter-métiers
  Soizick Lesteven (Observatoire Astronomique de Strasbourg)
- Pérennisation et interopérabilité des données de l'Observatoire de Recherche Méditerranéen de l'Environnement
  Juliette Fabre et **Olivier Lobry** (Observatoire des Sciences de l'Univers - OREME)
- Make your data great (again)
  Daniel Jacob (Institut National de la Recherche Agronomique - Inra)
- Open Data pour la simulation et les expérimentations au LEGI
  Gabriel Moreau, **Antoine Mathieu**, **Cyrille Bonamy**, **Joël Sommeria**, **Julien Chauchat** (Laboratoire des Écoulements Géophysiques et Industriels)

Observatoire Virtuel et FAIR
Des principes fondamentaux à la pratique

André Schaaff
Observatoire Astronomique de Strasbourg

Laurent Bourgès
Observatoire des Sciences de l'Univers de Grenoble

Karin Dassas
Institut d'Astrophysique Spatiale, Bures-sur-Yvette

Jean-Michel Glorian
Institut de Recherche en Astrophysique et Planétologie, Toulouse

Jean-Charles Meunier
Laboratoire d'Astrophysique de Marseille

Michèle Sanguillon
Laboratoire Univers et Particules de Montpellier

Pierre Le Sidaner
Observatoire de Paris

L'Observatoire Virtuel France est une action spécifique du CNRS-INSU initiée en 2004 pour mettre en place et soutenir la participation française à l'International Virtual Observatory Alliance (IVOA) qui est un consortium travaillant, « à la W3C », à l'élaboration de protocoles et standards d'interopérabilité autour des données. Sa réunion annuelle est prolongée d'une journée et demi d'échanges techniques entre les personnes travaillant dans les services mettant à disposition des données dans les domaines astronomie, planétologie, physique de l'héliosphère (domaines couverts par la Section 17 du Comité National). Ce réseau technique s'est développé, associant des acteurs techniques du Virtual Atomic and Molecular Database Center et maintenant de la Research Data Alliance. Ses membres participent aux travaux de l'IVOA et les mettent en pratique dans leur centre de données ou laboratoire. Les principes de FAIR y sont implicitement présents depuis le début.

Pour le principe « trouvable », on peut citer le Registry qui constitue les pages jaunes de l'OV ou les UCD (Unified Content Descriptors) qui permettent d'unifier les noms de colonnes des milliers de tables présentes en astronomie. Un effort particulier est fait au niveau des métadonnées et de la préservation en général.

Pour le principe « accessible », les données sont rapidement ouvertes à la communauté, avec des règles souples, à quelques rares exceptions près.

Le principe « interopérable » est le fondement de l'écosystème des protocoles et standards développés depuis 2001 par l'IVOA. L'interopérabilité vise en priorité notre domaine mais les standards ouverts existants sont les bases de notre travail, et permettent une ouverture vers d'autres domaines. Un service comme B2FIND, développé par le projet européen EUDAT, peut « harvester » le Registry IVOA *via* OAI-PMH. À contre-exemple un protocole comme VOSpace (accès au stockage) est trop « domaine spécifique » et ne va pas dans le sens FAIR de l'interopérabilité.

Enfin, pour le principe « réutilisable », un standard en cours d'élaboration concerne la « Provenance ». Basée sur les travaux du W3C, c'est une brique indispensable à la réutilisation des données produites par des instruments, acteurs, etc. La description (les UCD en sont un exemple pertinent), la documentation des données sont incontournables.

La mise en pratique de ces principes a beaucoup d'aspects positifs, une visibilité / interopérabilité rapide pour les nouveaux venus, le croisement et la comparaison des données, l'interopérabilité également des outils.

Parmi les points « négatifs », on peut citer le redimensionnement éventuel de l'infrastructure matérielle pour absorber l'augmentation des accès, l'ajout de couches supplémentaires (protocoles) aux données. Des problèmes plus complexes comme l'utilisation abusive des données (aspiration de grandes quantités de données) sont plus difficiles à gérer. La citation des données est souvent problématique, mais l'utilisation des DOI améliore la situation.

Nous souhaitons présenter ce travail et échanger avec les autres réseaux autour des bonnes pratiques.

Les documentalistes et la FAIRisation des données scientifiques : un travail d'équipe inter-métiers

Soizick Lesteven
CDS-Université de Strasbourg, CNRS, (Observatoire astronomique de Strasbourg)

Depuis sa création en 1972, le Centre de Données astronomiques de Strasbourg (CDS) a pour mission de collecter, répertorier, améliorer et distribuer les données astronomiques à la communauté astronomique internationale *via* les services SIMBAD (9,4 millions d'objets astronomiques), VizieR (17000 catalogues) et Aladin (atlas interactif du ciel). Sur l'ensemble des nos services nous comptabilisons un million de requêtes par jour.

Les différents services se sont construits et enrichis au fil du temps en fonction des besoins des utilisateurs et de l'évolution technique. Cette évolution est guidée par les besoins de la Science. Ces services reposent sur une « équipe intégrée » de documentalistes qui ont en charge la création du contenu des bases de données, d'informaticiens qui créent les systèmes et développent les outils et de chercheurs qui font évoluer les fonctionalités des services et le contenu des bases en fonction des besoins de la Science. La complémentarité des différents métiers et les échanges permanents entre ces différents acteurs est un des facteurs de notre réussite. Le travail des documentalistes évolue continument depuis la création du CDS, avec les évolutions scientifiques et technologiques.

Il est important de noter que le travail du CDS n'est possible que grâce à nos implications dans différents projets, collaborations et à l'aide de nos différents partenaires.

La logique FAIR est présente dans nos services depuis le début du CDS. Les documentalistes sont responsables de l'ingestion des données qu'ils/qu'elles vont tout d'abord identifier, sélectionner, vérifier, avant de les comparer, de les décrire et de les homogénéiser. Des interactions avec les chercheurs de l'équipe sont nécessaires pour résoudre ou valider des cas complexes. Cette forte valeur ajoutée aux données a un impact important sur la recherche car les données bien décrites et validées sont immédiatement réutilisables. Nos services sont aussi fortement utilisés car interopérables avec les autres centres de données en astronomie et grâce aux outils et aux standards de l'observatoire virtuel (VO) développés par les informaticiens et les astronomes. Notre challenge est de gérer l'augmentation des données tout en continuant à assurer la qualité des données distribuées et leur documentation.

Ce que nous souhaitons présenter ici est notre expérience de travail inter-métiers dans la gestion des données en astronomie qui pourrait être transposable dans d'autres domaines avec le développement de l'Open Science.

Pérennisation et interopérabilité des données de l'Observatoire de Recherche Méditerranéen de l'Environnement (OSU OREME)

Juliette Fabre, Olivier Lobry
OSU OREME, Montpellier

La pérennisation et la diffusion des données constituent une des missions principales de l'Observatoire de Recherche Méditerranéen de l'Environnement (OSU OREME[1]). Cet observatoire regroupe une vingtaine de Services d'Observation de huit laboratoires de recherche autour de l'environnement qui produisent une très grande diversité de données (données de capteurs, analyses génétiques, données cartographiques, photos, ...) sur des thématiques variées (hydrologie, géophysique, écologie, …).

Le service Système d'Information de l'Oreme développe les outils permettant la mise à disposition[2] des données d'observation dans le respect des normes et standards du domaine, et notamment en lien avec la directive européenne Inspire.

La grande diversité de types de données et de thématiques scientifiques soulève de nombreuses questions en matière de structuration en bases de données, d'utilisation de référentiels, de diffusion, d'interopérabilité, de catalogage, etc.

Pour pallier cette hétérogénéité, nous essayons d'avoir une approche la plus générique possible dans la constitution du système d'information, d'abord en terme de méthodes et d'interaction avec les producteurs de données, puis en matière d'outils, de standards et de normes, et enfin de développements.

La constitution du Système d'Information de l'Oreme s'appuie sur un dialogue constant et régulier avec les producteurs de données afin d'abord d'identifier leurs données d'intérêt (format, volume, standards du domaine, etc). Il s'agit ensuite d'identifier les traitements appliqués à ces données (filtres, corrections, validation, …), et enfin les besoins en terme de diffusion (type de diffusion, conditions d'utilisation, etc.).

Les données brutes sont structurées dans des bases de données relationnelles. Ces bases contiennent également les niveaux de données supérieurs (données corrigées, calculées, validées, …) ainsi que la description des processus permettant de passer d'un niveau n à un niveau n+1.

À ces données sont associés en base de données de nombreux descripteurs permettant de préciser finement le contexte des données en répondant aux questions qui, quoi, où, quand, comment et pourquoi.

Le fait de stocker les différents niveaux de données et l'ensemble du contexte permet d'aider à la ré-utilisabilité des données.

Pour assurer l'interopérabilité des données, nous intégrons autant que possible des référentiels dans les données et leurs descriptions : référentiels taxonomiques[3] et géographiques[4], thésaurus environnementaux et de biodiversité[5] et vocabulaires contrôlés[6]. Cette approche nous permet de déduire de nouvelles informations (par exemple la classification taxonomique d'un taxon ou l'arbre de confluence d'un cours d'eau) et ainsi d'enrichir automatiquement les données et de permettre de les lier entre elles.

---

[1] http://www.oreme.org

[2] http://data.oreme.org

[3] Catalogue of Life (http://www.catalogueoflife.org/), Taxref (https://inpn.mnhn.fr/telechargement/referentielEspece/referentielTaxo)

[4] BD Carthage et BD Geofla de l'IGN

[5] GEMET (http://www.eionet.europa.eu/gemet/en/about/), EnvThes (http://vocabs.ceh.ac.uk), Agrovoc (http://aims.fao.org/vest-registry/vocabularies/agrovoc), T-Semandiv (https://www.loterre.fr/skosmos/BLH/fr/) à venir

[6] GCMD (https://gcmd.nasa.gov/learn/keyword_list.html) pour les instruments, EBV (https://geobon.org/ebvs/what-are-ebvs/) pour les variables en biodiversité

Les données sont cataloguées selon le standard CSW de diffusion des métadonnées de l'OGC[7]. Le catalogue est généré automatiquement[8] à partir des données et des descriptions stockées en base. Il pourra être moissonné par des portails nationaux ou internationaux pour permettre aux données d'être plus facilement trouvables.

Pour assurer l'accessibilité des données, les fiches de métadonnées mentionnent entre autres (i) les conditions d'utilisation des données (licence Creative Commons CC-BY, à défaut de la Licence Ouverte qui ne dispose pas de version anglaise pour le moment), (ii) les éventuels identifiants uniques de données (DOI), (iii) ainsi que les URL d'accès aux données cartographiques (diffusées selon les standards WMS et WFS de l'OGC) et numériques (pages web permettant l'export des données en CSV).

Nous prévoyons de diffuser les données numériques selon le standard SOS de l'OGC, bien que les outils associés nous semblent à l'heure actuelle plus complexes à mettre en œuvre que ceux existants pour les autres standards de l'OGC[9].

À terme il serait également intéressant d'utiliser les technologies du web sémantique pour diffuser les données et les lier automatiquement avec d'autres ressources en exploitant les référentiels utilisés dans notre base.

Des DOI[10] peuvent enfin être générés automatiquement sur nos jeux de données. Toutefois, le respect des règles imposées à un jeu de données référencé dans Datacite requiert une certaine réflexion et stratégie pour des bases de données dynamiques comme la nôtre. Il faut ainsi prendre en compte les modifications de données (correction, suppression) et leur versionnement, pour que les données puissent être retrouvées dans l'état exact dans lequel elles ont été citées.

Rendre les données FAIR dans le cadre d'un Observatoire de l'Environnement implique donc de maîtriser tout un écosystème de compétences (conception de bases de données, développement, SIG, web sémantique, etc.) ainsi que les concepts, outils et technologies associées, qui peuvent présenter des degrés de maturité ou des facilités d'utilisation relativement différentes.

---

7 http://www.opengeospatial.org/standards

8 librairies R geometa (https://cran.r-project.org/web/packages/geometa/index.html) et geonapi (https://cran.r-project.org/web/packages/geonapi/index.html)

9 Ex Geoserver pour le WMS/WFS, et GeoNetwork pour le CSW

10 Digital Object Identifier

Make your data great (again)

Daniel Jacob
Institut national de recherche agronomique-INRA
Biologie du Fruit et Pathologie (UMR 1332), Villenave D'Ornon

Les données scientifiques ont de la valeur au-delà de la question de recherche initialement posée. Pour qu'elles aient une deuxième vie en quelque sorte, il faut que l'ensemble des acteurs dans la chaîne d'acquisition de ces données soient convaincus que le dépôt des données peut apporter de la valeur ajoutée et ce, d'autant plus que les producteurs le feront le plus tôt possible, par exemple l'aide à la prise de décision dans la sélection d'échantillons, dans le choix d'analyses complémentaires, aide à l'annotation par croisement, apport de connaissance *a priori*, etc. La « vie de la donnée » doit donc être intégrée dans le processus de la recherche scientifique obligeant les (bio) informaticiens à proposer des outils utiles et/ou innovants pour motiver et convaincre les chercheurs de déposer leurs données le plus tôt possible. Notre expérience dans la manipulation de données, essentiellement phénotypiques (physiologie et métabolisme des plantes) issues d'expérimentations en serres ou en champs, a permis de mener une réflexion sur l'optimisation de la gestion du flux de données. Deux préoccupations majeures ont été à l'origine de cette réflexion : (i) la capture des données et (ii) leur utilisation le plus tôt possible par un ensemble de partenaires multi-sites. Notre analyse nous a mis en évidence la nécessité de rendre cohérent le flux de données (data flow) proprement dit d'une part (à savoir du design expérimental jusqu'à la diffusion des données en passant par leur mise en forme et leur annotation) et d'autre part les étapes d'analyses de données (data process : data mining & modelisation). Nous avons développé un framework (ODAM Open Data, Access and Mining)[1] permettant le plus simplement possible de rendre les tableaux de données expérimentales largement accessibles et entièrement réutilisables, y compris par un langage de script tel que R. Le but principal est de rendre un ensemble contextualisé de données accessible en ligne avec un minimum d'effort de la part du fournisseur de données. L'approche consiste à construire un réseau de données sur le web, basée sur des technologies appropriées (Web API, Linked Data), et en utilisant les formats standards de données (TSV, JSON, XML, RDF). Des applications web ayant chacune un objectif bien délimité, viennent ensuite exploiter ce réseau. Une donnée peut donc servir à plusieurs applications et vice-versa. Le système de gestion de la donnée devient totalement indépendant de son exploitation. La donnée est ainsi « décloisonnée », condition *sine qua non* pour le Web of Data.

[1] https://fr.slideshare.net/danieljacob771282/data-management-djinra2018/

Open Data pour la simulation et les expérimentations au LEGI

Gabriel Moreau, Antoine Mathieu, Cyrille Bonamy, Joël Sommeria, Julien Chauchat
Laboratoire des Écoulements Géophysiques et Industriels-LEGI (UMR5519), Grenoble

Le LEGI s'est engagé depuis de nombreuses années dans la voie de l'Open Data. À l'origine, il y a une dizaine d'années, il s'agissait de partager le résultat de simulations numériques de dynamique océanique à l'aide d'un serveur interne DAP. Avec l'explosion de la pratique de l'OpenScience, mettre à disposition ses données à la communauté s'est généralisée au sein du laboratoire.

Le LEGI est fortement impliqué dans la problématique de l'Open Data au travers des données obtenues sur la plateforme CORIOLIS, l'une des grandes installations expérimentales du projet européen Hydralab+ (2015-2019). Des chercheurs du monde entier viennent y réaliser des expériences et la Commission européenne impose la diffusion libre des données acquises dans le cadre de ces programmes après une période d'embargo de deux ans. Dans le cadre d'Hydralab+ un Joint Research Activity (JRA) est dédié à ces aspects Open Data.

Les données acquises sur la plateforme CORIOLIS sont principalement des images et des champs de vitesse organisés en dossiers par numéro d'expérience et nom de caméra, sauvés sous des formats standards, pérennes, multi-plateformes et multi-langages (PNG, NetCDF, XML). Cette organisation des données permet à une personne extérieure au projet initial de reprendre des anciennes données et de les analyser. Un point important pour la compréhension des données est d'associer à chaque projet une page web (Wiki), décrivant le cadre des expériences ainsi que le jeu de paramètres de chaque essai.

Dans le cadre de simulations numériques sur le transport sédimentaire, la démarche complète de l'Open Data a été accomplie. Les données ayant permis la réalisation des courbes et des tableaux d'une publication ont été téléversées sur Zenodo sous licence libre. Le choix des données pertinentes et leurs mises en forme sous format pérenne, l'explication de l'organisation des données, ont permis de développer un certain savoir-faire que nous nous efforçons de généraliser aux données issues de la plateforme CORIOLIS.

Les archives Zenodo sont effectivement une solution pour l'archivage des données associées à une publication et permettent à une personne de reprendre les données et de reproduire les courbes. Afin de rendre la construction des archives Zenodo plus robuste et plus facile pour les utilisateurs, nous avons développé un fichier de meta-données (PROJECT-META.yml) de type clef / valeur permettant de vérifier la présence de clefs et ainsi de générer automatiquement un certain nombre de fichiers (README, AUTHOR, LICENCE…) pour finalement créer la ou les archives Zenodo associées.

Cependant, les archives Zenodo sont limitées en termes de taille (50 Go) et d'accès (téléchargement complet de l'archive). Notre serveur DAP (en pratique une instance OpeNDAP) possède l'avantage de ne pas imposer de quota et de permettre le téléchargement partiel des données ou une lecture directe en ligne.

Enfin, pour rendre les données diffusées plus compréhensibles et utilisables par la communauté nous avons développé des notebooks Python disponibles *via* Github et utilisables facilement avec Binder (un Jupyter ouvert à tous). Ainsi, grâce à un outil visuel non hébergé au LEGI (Jupyter / Binder / Python), il est possible de visualiser les données Open Data du LEGI sans aucun téléchargement sur son poste de travail.

Partie I - Session 4
Interventions du 27 novembre

Synthèse de la journée

- Présentation de RDA France - **Francis André** (Direction de l'Information Scientifique et Technique du CNRS)
- Conclusion synthèse journée[1] - **Volker Beckmann** (Directeur Adjoint Scientifique en charge du domaine « Calcul et données » de l'Institut national de physique nucléaire et de physique des particules - IN2P3)

Discussion
- Échanges avec les participants[2]

[1] L'intervention de Volker Beckmann n'ayant pas fait l'objet d'une retranscription, vous pouvez la retrouver sur la captation *Synthèse de la journée interopérabilité et pérennisation des données de la recherche* à partir de : 19'20"

[2] Les échanges avec les participants n'ayant pas fait l'objet d'une retranscription, vous pouvez les retrouver sur la captation *Synthèse de la journée interopérabilité et pérennisation des données de la recherche* à partir de : 33'20"

Présentation de RDA France

Francis André
Direction de l'Information Scientifique et Technique du CNRS)

La Research Data Alliance - le Nœud National RDA France

L'Alliance pour les données de recherche (RDA - Research Data Alliance) est une organisation internationale qui œuvre à faciliter le partage des données de recherche en bâtissant des passerelles sociales et technologiques par-delà métiers et disciplines scientifiques. RDA est une organisation pilotée par ses membres ; elle s'appuie sur une communauté de plus de 8100 personnes provenant de 137 pays différents, couvrant large spectre de métiers et de communautés thématiques. Elle offre un forum où les membres se réunissent pour développer et adopter des solutions qui favorisent la recherche et le partage de données et accélèrent la croissance d'une communauté de données. Une grande variété de sujets ayant trait aux données de recherche sont travaillés dans plus de 90 groupes de travail et groupes d'intérêts.

La RDA est soutenue financièrement par les États-Unis, l'Australie et la Commission européenne. Cette dernière soutient les activités de la RDA à travers une série de projets « RDA Europe » depuis le lancement de l'Alliance en 2013.

Le Nœud national de la RDA : animer une communauté d'experts

RDA Europe 4.0, le projet européen actuel en soutien des activités de l'Alliance, s'est donné pour objectif de développer des nœuds nationaux, relais indispensables pour irriguer les communautés scientifiques nationales sur ces thématiques du partage des données, compte tenu de la diversité de structuration de la recherche dans les différents pays européens.

Le CNRS est en charge de développer le nœud national RDA français qui est cité dans le Plan National pour la Science Ouverte et travaille en lien avec le Collège Données du Comité pour la Science Ouverte. L'objectif principal de RDA France n'est pas de créer une plate-forme uniquement pour l'échange de connaissances, mais plutôt d'opter pour des approches plus pratiques, en lançant des projets d'adoption concrets pour diffuser les résultats.

Le nœud national met l'accent sur les actions suivantes :

- le développement et animation d'une communauté RDA nationale ;
- la promotion des activités de l'alliance (appels à propositions, conférences plénières, etc.) en facilitant la participation de membres français aux différents groupes de travail et groupes d'intérêt ;
- l'organisation et ou participation à des évènements (réunions, ateliers, sessions de formation, etc.) pour présenter et stimuler l'adoption des résultats de la RDA ;
- les interactions avec les acteurs nationaux de la recherche : communautés scientifiques, agences de financement, organismes de recherche, universités, etc.

Partie II

Autres propositions d'interventions

(par ordre alphabétique)

- Diffuser des données d'enquêtes en SHS. Le FAIR data comme outil d'harmonisation
  Valentin Brunel, Sarah Cadorel (Centre de Données Socio-Politique – CDSP, Sciences-Po, Paris)
- Le système d'information commun de données in situ OZCAR - Theia
  Véronique Chaffard, Charly Coussot, Sylvie Galle, Patrick Juen (Université Grenoble Alpes, CNRS, IRD, Grenoble-INP, IGE, Grenoble), **Isabelle Braud** (Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture - Irstea, Villeurbanne)
- ArchiPolis - Référencement, indexation et sensibilisation à l'archivage raisonné des enquêtes en sciences sociales du politique
  Émilie Fromont (Centre de Données Socio-Politique – CDSP, Sciences-Po, Paris)
- Mise en place d'un plan de gestion des données au GANIL
  Benjamin Fusberti (Grand accélérateur national d'ions lourds - GANIL - CNRS - UPR3266, Caen)
- Du Plan de Gestion des Données au datapaper. Gestion durable des données scientifiques tout au long de leur cycle de vie
  Wilfried Heintz (Institut national de recherche agronomique - INRA - Dynafor, Castanet Tolosan)
- Le Système d'Information Nature et Paysages
  Rémi Jomier (UMS Patrinat - MNHN/AFB/CNRS, Paris)
- Pôle National de données de biodiversité
  Yvan Le Bras (Station Martine de Concarneau - MNHN PatriNat - UMS2002, Concarneau)
- Retour d'expériences et bonnes pratiques des données sur l'eau transposables aux données de la recherche
  Nathalie Moulard (Direction générale de l'agence de l'eau Loire-Bretagne, Gouvernance des données - CNRS, Orléans)
- Corpus factory
  Michaël Nauge (Formes et représentations en linguistique et littérature - FoReLLIS/ Mémoires, Identités, Marginalités dans le Monde Occidental Contemporain-MIMMOC - Université de Poitiers), **David Chesnet** (Maison des Sciences de l'Homme et de la Société - MSHS, Poitiers)

- Le Centre de Données de Géothermie Profonde. Un exemple FAIR
  Marc Schaming, Jean Schmittbuhl (Institut de physique du globe de Strasbourg - IPGS - UMR 7516, Université de Strasbourg, CNRS, Strasbourg), **Alice Frémand, Marc Grunberg** (École et observatoire des sciences de la Terre-EOST - UMS830, Université de Strasbourg, CNRS, France)
  Nicolas Cuenot (Électricité de Strasbourg-Géothermie, Strasbourg)

- SSHADE - L'infrastructure Européenne de bases de données en spectroscopie et spectro-photométrie des solides
  Bernard Schmitt, Philippe Bollard, Alexandre Garenne, Damien Albert, Lydie Bonal (IPAG, OSUG - DC / UGA - CNRS, Grenoble), **et les partenaires du consortium SSHADE**

- dat@OSU, une plateforme de référencement et de valorisation des données de la recherche à l'OSU THETA
  Hélène Tisserand (OSU THETA de Franche-Comté/Bourgogne - UMS3245 - Observatoire de Besançon), **Sylvie Damy** (Laboratoire Chrono-environnement - UMR6249), **Bernard Debray** (Institut UTINAM - UMR6313), **Raphaël Melior** (OSU THETA de Franche-Comté Bourgogne - UMS3245)

Diffuser des données d'enquêtes en SHS
Le FAIR data comme outil d'harmonisation

Valentin Brunel, Sarah Cadorel
Centre de Données Socio-Politique–CDSP - Sciences-Po, Paris

En 2006, le CDSP prend la suite de la BDSP (Banque de Données Socio-Politiques) dans sa mission de documentation et de diffusion d'enquêtes en sciences sociales.

À ce titre le CDSP participe encore à un réseau de diffuseurs (le réseau Quételet) qui comprend aussi des organismes de diffusion de la statistique publique. Pendant quelques années le CDSP a ainsi mis en place des procédures unifiées plus ou moins normées pour la diffusion de données en SHS.

Parmi les réalisations de cette première phase d'expansion du CDSP, on peut compter le développement du portail Quételet, qui servira désormais d'application de commande des données diffusées, ainsi que la création d'une base de questions visant à faciliter la navigation au sein des données (recherche par mots-clefs, par diffuseur, par vocabulaire, etc.).

En 2012, le CDSP voit débuter l'Equipex DIME-SHS, qui est l'occasion d'un développement important des activités du centre. À l'occasion de cet octroi, le CDSP devient producteur de données avec le panel web ELIPSS et documente désormais des enquêtes qualitatives (par entretien, observation...) dans la banque d'enquêtes beQuali. Cette diversification entraîne l'application de plusieurs principes du FAIR Data : normalisation des langages de métadonnées utilisés - due notamment à des interactions de plus en plus nombreuses avec des prestataires des organisations ou des chercheurs (facilitant l'interopérabilité) - surcroît de communication autour des données et de l'activité du centre, mais aussi plus grande attention portée à la réutilisation des données par la communauté scientifique, tant du point de vue des moyens que des réalisations, du fait du regard attentif des bailleurs de fonds.

Depuis 2012, c'est donc un véritable tournant vers le FAIR Data qui caractérise l'évolution des pratiques du CDSP.

Les pratiques de diffusion et de documentation issues d'univers variés ont convergé à travers trois chantiers principaux : le début du traitement de données « quanti-quali », la création de solutions techniques adaptées aux enquêtes et l'archivage pérenne des enquêtes diffusées.

Le traitement de données d'enquêtes « quanti-quali » (par questionnaire et par entretien) a permis une homogénéisation des métadonnées DDI pour décrire l'étude et son contexte. Par la même occasion le schéma de description a été éclairci et normalisé. De même, afin de faciliter l'interface avec les utilisateurs, les contrats d'utilisation des données ont été harmonisés, ces dernières restant accessibles à travers un portail unique d'accès.

La création de solutions techniques adaptées aux enquêtes a permis d'avancer sur les questions de la traçabilité et de la réutilisation des données. Parmi ces solutions, les outils d'exploration des corpus et des variables visent à faciliter la réutilisation mais aussi améliorer l'accessibilité des métadonnées, aujourd'hui disponibles pour tous.

Enfin le chantier d'archivage pérenne a été l'occasion de travailler à la conservation des données diffusées et de leur documentation. Les formats de données publiés ont aussi été adaptés pour correspondre à l'exigence d'ouverture lors de l'archivage (pas de format propriétaire). Plus généralement l'archivage a été l'occasion de voir l'activité de documentation s'étendre au cycle entier de la vie des données.

Le système d'information commun de données *in situ* OZCAR - Theia

Véronique Chaffard, Charly Coussot, Sylvie Galle, Patrick Juen
Université Grenoble Alpes, CNRS, IRD, Grenoble-INP, IGE, Grenoble

Isabelle Braud
Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture-Irstea, Villeurbanne

L'infrastructure de Recherche (IR) sur la zone critique OZCAR (Observatoires de la Zone Critique Applications et Recherches), futur miroir français de l'IR européenne e-LTER *european Long Term Ecological Research* en cours de construction, et le pôle national de données des surfaces continentales Theia sont en train de construire un système d'information (SI) commun qui a pour objectif de mettre en visibilité l'ensemble des données *in situ* d'observation des surfaces continentales et d'en faciliter l'accès sur un portail unique.

Les enjeux dans la construction de ce SI sont de faire du porter à connaissance des données d'observation de la zone critique, de faciliter leur recherche et leur exploration, de permettre leur pérennisation et leur citation, de favoriser leur réutilisation et leur partage. Le système devra être interopérable avec les infrastructures de données européennes en cours de construction.

Pour servir ces enjeux, le rôle des métadonnées accompagnant les données est central et la mise en œuvre de l'interopérabilité impose une standardisation de ces métadonnées à la fois sémantique et technique.

Les premières étapes dans la construction du SI ont donc consisté à :

- identifier les différents standards de métadonnées existants (ISO 19115/Inspire, DataCite, DCAT, schema.org, SensorML, O&M) pour couvrir les besoins applicatifs et d'interopérabilité et établir un set minimum de métadonnées à échanger avec les fournisseurs ;
- définir un format pivot pouvant véhiculer l'ensemble des métadonnées identifiées ;
- identifier les thesaurii internationaux pertinents pour le domaine et publiés sur le Web ;
- construire et publier sur le Web un vocabulaire contrôlé pour les noms de variables servant de critère de recherche de la donnée, en s'appuyant sur les thesaurii identifiés et en liant les concepts *via* des relations sémantiques SKOS.

La présentation s'attachera à exposer la démarche susceptible d'être réutilisée dans d'autres disciplines.

ArchiPolis - Référencement, indexation et sensibilisation à l'archivage raisonné des enquêtes en sciences sociales du politique[1]

Émilie Fromont
Centre de Données Socio-Politique–CDSP, Sciences-Po, Paris

Le projet ArchiPolis, consortium de l'Infrastructure de Recherche Corpus (Huma-Num 2012-2016), fédère des laboratoires de sociologie et de science politique au sein desquels sont ou ont été menées des recherches qualitatives portant sur l'objet politique au sens large. Le projet continue actuellement sous forme de réseau coopératif.

Les objectifs étaient de développer des procédures et des standards numériques partagés pour la préservation des données de la recherche des enquêtes par entretiens menées au sein des laboratoires. Nous voulions également sensibiliser les chercheurs à la question de la conservation, de l'archivage, de l'identification de leurs données et enfin réaliser un inventaire des enquêtes qualitatives produites par chaque structure.

Les corpus recouvrent plusieurs champs des sciences sociales du politique (sociologie, science politique, ethnologie, etc.) et sont constitués de matériaux qualitatifs diversifiés : entretiens (dont certains ont été enregistrés), observations de terrain, grilles d'analyse ou documents préparatoires.

Dans le but de valoriser les matériaux d'enquêtes autant que les publications effectuées à partir des travaux de terrain, le consortium a accompli un inventaire exhaustif des enquêtes réalisées ou en cours de réalisation dans les laboratoires partenaires. La mise en ligne de cet inventaire s'inscrit pleinement dans l'esprit FAIR.

L'outil OpenSource Dataverse, développé par l'université de Harvard, a été adopté pour partager et diffuser ces informations, ce qui a amené le consortium à travailler sur l'homogénéisation, la sélection et l'interopérabilité des métadonnées partagées. Le Dataverse permet de décrire les enquêtes par des métadonnées standardisées offrant une contextualisation riche (titre, dates extrêmes, auteurs, méthodologie suivie, principales publications, lieu de conservation des documents, etc.). L'outil intègre en effet les standards internationaux de description d'enquête de la Data Documentation Initiative (DDI), le Dublin Core et certaines métadonnées de géolocalisation. Les descriptifs sont librement accessibles à tout internaute et les métadonnées moissonnables par le protocole OAI-PMH. Certaines métadonnées sont par ailleurs liées à d'autres ressources (publications, liens vers des données disponibles ailleurs, comme par exemple les enquêtes numérisées dans le cadre de BeQuali). Un identifiant pérenne (DOI) est associé à chaque enquête, facilitant la citation.

Le portail permet à chaque laboratoire de gérer et de valoriser son propre catalogue, tout en rassemblant les enquêtes de tous les laboratoires au sein du **Dataverse ArchiPolis** qui contient actuellement plus de 220 notices d'enquêtes. Celles-ci peuvent être mises à jour et un versionning permet de suivre l'historique des modifications.

Le portail est hébergé par Sciences Po à Paris : **https://catalogues.cdsp.sciences- po.fr/dataverse/archipolis**, par l'intermédiaire du Centre de données socio-politiques, membre du réseau.

Le projet a par ailleurs permis de traiter un grand nombre de problématiques induites par cet archivage raisonné : attachement des chercheurs à des données qu'ils s'approprient, nécessité de sensibiliser les doctorants, articulation avec les nouvelles pratiques de plans de gestion des données, traitement de données sensibles et anonymisation, conditions de conservation de supports vulnérables ou obsolètes, conditions de conservation pérenne *via* des services d'archivage, mise en place de la collaboration entre des chercheurs et des personnels d'appui à la recherche (archivistes, documentalistes, ingénieurs méthode ou informaticiens).

[1] consortium labellisé Huma-Num de 2012 à 2016. Membres d'ArchiPolis : le Centre de données socio-politiques, le Centre d'études européennes, le CERAPS, le Centre de sociologie des organisations, Pacte, le Centre de recherches politiques de Sciences Po, l'Observatoire sociologique du changement, Triangle et le Centre Émile Durkheim (jusqu'en 2017)

Mise en place d'un plan de gestion des données au GANIL

Benjamin Fusberti
Grand accélérateur national d'ions lourds-GANIL (CNRS - UPR3266), Caen

Le projet

Une réflexion sur la gestion des données a été lancée, au GANIL, dans le cadre du projet européen IDEAAL (n°730989).

Le GANIL, Grand Accélérateur National d'Ions Lourds, est aujourd'hui l'un des grands laboratoires internationaux pour la recherche avec des faisceaux d'ions.

L'objectif du projet IDEAAL est d'explorer les possibilités de développement de l'infrastructure afin de garantir la viabilité du laboratoire à long terme.

L'objectif que nous nous sommes fixé en termes de gestion des données est de garantir la pérennité des données scientifiques produites au GANIL, de permettre un accès aux données à toute personne manifestant un intérêt pour celles-ci et d'augmenter la visibilité du laboratoire, des physiciens et des expériences.

Nos choix d'architecture fonctionnelle et technique se portent actuellement sur des outils simples et modulables avec pour objectif principal de répondre aux principes FAIR.

Problématiques traitées

Estimation des volumes de données actuels et à venir	Environ 200To aujourd'hui/ à 800To d'ici 3 ans
Cadrage des données à conserver	Données brutes / Configuration électronique de l'expérience / Documentations diverses (logbook, compte-rendu, documentation liée à la proposition de l'expérience) / Données réduites /Package logiciel
Métadonnées	Schéma fourni par DataCite, enrichi de métadonnées caractéristiques des expériences menées au GANIL
Mise en place d'un identifiant pérenne par expérience	Digital Object Identifier géré par DataCite
Choix d'une licence pour les données	Creative Commons (CC-BY 4.0)
Description du cycle de vie des données	De la création à la destruction éventuelle des données / Différents scénarii en fonction du type de recherche (publique, privée, collaboration)
Réflexion autour de l'Open Access	Accès à la demande après une période d'embargo de 3 ans prolongeable.
Réflexion autour de la création d'un portail des données	Étude en cours
Choix d'un entrepôt de données	CC-IN2P3
Transfert des données vers cet entrepôt	Effectué avec l'application iRODS

Rédaction d'une documentation liée aux données

Politique des données	Description de la propriété, de la responsabilité, de la gestion et de l'accès aux données
Plan de gestion des données du laboratoire	Description du cycle de vie des données expérimentales, et des process liés à leur gestion
Plan de gestion des données pour chaque expérience	Description de la gestion du jeu de données d'une expérience / Machine actionnable pour faciliter la gestion des données une fois stockées dans l'entrepôt

Communautés scientifiques concernées

Physiciens, astrophysiciens, radiobiologistes, …

Maturité des problèmes posés

L'objectif est de pouvoir mettre en place les outils destinés à une gestion FAIR des données pour le prochain RUN expérimental qui commence en avril 2019.

La majorité des problèmes a donc été traitée. Reste la mise en place technique et la rédaction des documents.

Implication des chercheurs

Pas facile à obtenir…

Entretiens en face à face réalisés avec la majorité des physiciens GANIL.

5 physiciens GANIL et 2 physiciens IN2P3 ont participé à différents workshops (métadonnées, données à préserver).

Problèmes rencontrés

Manque d'intérêt des chercheurs pour les problématiques de gestion des données.

—> Échange avec d'autres laboratoires ayant lancés ce type de démarche.

—> Échange ou informations avec DIST, INIST, RDA, IN2P3.

Changement des pratiques pour les chercheurs (utilisation d'un ORCID, enrichissement des métadonnées, ajout du DOI des données dans la future publication, …).

Démarches contraignantes mais impératives pour un management des données efficace.

—> Un accompagnement sera nécessaire.

Du Plan de Gestion des Données au datapaper Gestion durable des données scientifiques tout au long de leur cycle de vie

Wilfried Heintz
Institut national de recherche agronomique - INRA (Dynafor), Castanet Tolosan

cycle de vie des données, plan de gestion de données, infrastructure de données géographiques, curation de données, carnet de terrain électronique, R, Opidor, protocole, qualité des données, open science, métadonnées

Suite au constat de son incapacité à avoir une vue générale de ses données, indépendamment de leur ancienneté, l'UMR Dynafor s'est fixé pour objectif d'améliorer la gestion de ses données. Il s'agit en particulier de mieux partager les données que nous produisons, entre tous les agents de l'unité dans un premier temps, et avec le reste de la communauté à plus long terme.

Pour répondre à ces besoins, nos réflexions se sont basées sur la notion de cycle de vie des données et se sont nourries des avancées technologiques et méthodologiques amenées par le mouvement de l'Open Science.

La gestion durable des données commence dès la conception du projet pour lequel elles seront récoltées : grâce à un plan de gestion des données (PGD) partagé, tous les acteurs d'un projet auront à leur disposition les éléments essentiels concernant à la fois les données et leurs caractéristiques (nature, format, point d'accès, etc.) mais aussi les agents impliqués à un ou plusieurs moments de leur gestion.

Nous avons initialement porté nos efforts sur la mise en œuvre d'une infrastructure de données géographiques (IDG), afin de garantir une structure pérenne et interopérable pour nos données. Une IDG permet de coupler *a minima* un système de gestion de base de données (SGBD), un outil de visualisation cartographique et un catalogue des métadonnées.

Ces métadonnées constituent le fil conducteur de la gestion des données : elles sont élaborées avant les données elles-mêmes, et sont complétées de manière participative par tous les acteurs d'un projet tout au long de son cycle de vie. Ce sont elles qui permettront la recherche, la découverte et l'exploitation des données ; elles sont également le constituant principal d'un datapaper. Pas de FAIR sans métadonnées !

Un point essentiel de cette gestion repose donc sur le caractère partagé de l'ensemble des éléments : les documents de cadrage, logiciels et méthodologies aussi bien que les données elles-mêmes. Ceci suppose l'emploi de supports accessibles (outils Web), de formats ouverts et standardisés, conformément aux directives et standards internationaux.

Ainsi, nous avons mis en place une chaîne d'outils informatiques afin d'assurer une gestion optimale des données tout au long de leur cycle de vie, depuis la planification de leur collecte jusqu'à la publication d'un datapaper.

Dans notre exposé, nous présenterons les différents outils pratiques que nous avons mis en place. Un focus sera fait sur les outils embarqués qui nous permettent depuis peu d'initier un flux intégré de données, dès la collecte de celles-ci.

Le Système d'Information Nature et Paysages

Rémi Jomier
UMS Patrinat - MNHN/AFB/CNRS, Paris

Le Système d'Information sur la Nature et les Paysages (SINP) a notamment pour objet de :

- structurer les connaissances sur la biodiversité (faune, flore, fonge) ;
- mettre à disposition ses connaissances ;
- afin d'en faciliter la mobilisation pour élaborer ou suivre les politiques publiques, évaluer les impacts des plans, programmes, projets des différents aménageurs ;
- et de permettre le rapportage correspondant aux engagements européens et internationaux.

Le SINP est un dispositif partenarial entre le Ministère chargé de l'environnement, les associations, les collectivités territoriales, les établissements publics et opérateurs, les services de l'Etats, etc.

Le SINP privilégie une organisation en réseaux et repose sur des producteurs de données, des plateformes régionales ou thématiques et une plateforme nationale. Tous adhèrent, *via* l'adhésion à un protocole, à un ensemble de règles et de principes communs.

Une des difficultés à l'échange d'informations dans ce système découle de la multitude d'acteurs qui produisent de la donnée : cette diversité implique des pratiques différentes, des formats extrêmement divers (allant du carnet de terrain papier à la base de données centralisée nationalement, en passant par la simple feuille de calcul), et un nombre d'informations renseignées variable en fonction des acteurs.

Le SINP propose donc des référentiels (TAXREF, HABREF, etc.) et des dictionnaires de données afin de faciliter l'interopérabilité. Il propose également des « mappings » (ou documents de correspondance) pour permettre de valoriser ces données au niveau international.

Des procédures de contrôle sont mises en place (conformité, cohérence et validation scientifique) afin d'augmenter la fiabilité des données diffusées et en permettre une meilleure réutilisation.

Le SINP met également en place une organisation permettant la description des données (*via* le renseignement de métadonnées). Cela permet également aux ré-utilisateurs de mieux sélectionner les données utiles pour leurs études. Cela permet également de mieux valoriser les différents acteurs de la donnée (maitre d'œuvre, ouvrage, financeur, etc.).

Toutes ces données sont diffusées et rendues accessibles. Une application nationale dédiée est en cours de construction.

Pôle National de données de biodiversité

Yvan Le Bras
Station Martine de Concarneau - MNHN PatriNat (UMS2002), Concarneau

Le 8 mars dernier, le Ministère de l'Enseignement supérieur, de la recherche et de l'innovation a inscrit sur sa feuille de route la création d'une nouvelle infrastructure intitulée Pôle National de données de biodiversité (PNDB).

Les missions du PNDB s'inscrivent dans une approche FAIR et consistent notamment à **1**. fournir un accès aux jeux de données et de métadonnées, à des services associés et à des produits dérivés des analyses ; **2**. favoriser la cohérence avec les efforts nationaux, européens et internationaux relatifs à l'accès et à l'exploitation des données de recherche sur la biodiversité, à la promotion de produits et services.

Pour atteindre ces objectifs, le PNDB se propose de mettre en place une e-infrastructure sous la forme d'un système open-source facilitant, pour les communautés en Écologie, l'accès aux et la gestion des données et métadonnées d'Écologie mais aussi des traitements analytiques. L'ambition FAIR assumée de l'infrastructure induit une réflexion profonde en cours sur la manière de pouvoir réellement concilier les pratiques actuelles du milieu de la recherche en Écologie avec un degré de FAIRness maximum. En partant d'un premier niveau de conceptualisation du cycle de vie de la recherche en Écologie, différentes briques open-sources identifiées sont actuellement éprouvées et concernent notamment :

Pour les aspects métadonnées :

- L'utilisation de l'EML comme langage pivot pour décrire et échanger les métadonnées des projets de recherche ;
- L'utilisation d'un catalogue de métadonnées de type Metacat pour mettre α disposition les métadonnées et pointer vers les données ;
- L'utilisation d'outils de saisies de métadonnées de type IPT-Toolkit.

Pour les aspects analyses :

- Le recours au packaging conda *via* Bioconda associé α de la containerisation *via* Biocontainer pour assurer un haut degré d'accessibilité et de reproductibilité des composants analytiques, majoritairement développés sous la forme de scripts R ;
- L'utilisation de la plateforme collaborative de consultation et analyse de données en Écologie Galaxy-E pour permettre un accès transparent aux données et composants analytiques et assurer une haute reproductibilité des anlayses effectuées grâce aux fonctionnalités d'historiques et de création, partage de workflows d'analyse.

Retour d'expériences et bonnes pratiques des données sur l'eau transposables aux données de la recherche

Nathalie Moulard
Direction générale de l'agence de l'eau Loire-Bretagne, Gouvernance des données » / CNRS, Orléans

L'agence de l'eau Loire-Bretagne dispose d'une politique des données, avec un poste dédié à la direction générale. Les données sont une ressource partagée, font partie de son patrimoine, concourent directement à la réussite de ses missions et à la qualité de son image de marque. Comme dans d'autres organismes publics des freins existent et sont progressivement levés. L'agence collecte, utilise et publie de nombreuses données, pour partie sur des plateformes nationales dont celles du Système d'Information sur l'Eau (SIE).

Plusieurs bonnes pratiques pourraient être transposées au monde de la recherche afin d'améliorer la réutilisation des données selon les principes FAIR :

F. Faciliter la découverte : Les données sont inventoriées ou moissonnées dans des catalogues ou géocatalogues (data.gouv.fr, géocatalogue.fr). Pour l'agence de l'eau, l'utilisation des données par des tiers a ainsi été multipliée par 2 000.

A. Rendre accessible : La plateforme du Sandre, pour le SIE, permet un accès unitaire à chaque donnée par une URL stable, standardisée et pérenne. Des liens directs vers ces URI sont alors possibles. L'implémentation d'URL contextuelles pour des valorisations (telles que la BNPE) permet de construire des moteurs de recherche thématiques ou territoriaux, d'inclure des liens vers des données valorisées dans des articles pour des non experts…

I. Rendre interopérables les données : Les web services cartographiques (OGC : WFS et WMS) permettent aux cartographes d'utiliser les données, en ligne, sans les télécharger. L'agence de l'eau utilise la plateforme Carmen pour exposer ainsi ses données cartographiques. Les données non cartographiques peuvent être exposées par des web services, en suivant des standards nationaux (le Sandre décrit ces standards pour le SIE) ou internationaux. Hélas, ces standards sont souvent difficiles à implémenter pour le ré-utilisateur. Cela a conduit le SIE à construire la plateforme Hub'eau qui offre des API clé en main permettant d'utiliser plus simplement les données.

R. Faciliter la réutilisation des données : avec les données, il est préférable de diffuser :

- des liens avec des données de référence, ce qui facilite l'utilisation des données dans des systèmes existants où ces données de référence existent déjà ;
- d'informations sur le contexte d'acquisition de la donnée (où, quand, comment)… ;
- d'informations sur la validation (niveau de qualité des données) ;
- la documentation des données (dictionnaire quand il n'existe pas de standard, liste de valeurs avec leur signification…).

Corpus factory

Michaël Nauge
Formes et représentations en linguistique et littérature-FoReLLIS/ Mémoires, Identités, Marginalités dans le Monde Occidental Contemporain-MIMMOC - (Université de Poitiers),Poitiers

David Chesnet
Maison des Sciences de l'Homme et de la Société-MSHS, Poitiers

Les recherches scientifiques en sciences humaines et sociales sont très variées et semblent très différentes, cependant nombre d'entre elles ont la particularité de partager des processus et des méthodes pour les études sur corpus numériques.

Notre projet s'est fixé pour objectif de développer une chaîne générique de traitement automatisé facilitant la création de corpus numériques conformes aux principes FAIR sur l'infrastructure nationale Huma-Num. Notre but est de libérer un maximum de ressources aux chercheurs pour l'exploration et l'analyse des données en ayant optimisé le temps de production, la qualité et la réutilisabilité des corpus.

Pour espérer créer ce type de corpus, il est impératif de s'appuyer sur des infrastructures de confiances, pérennes et des outils fiables. Les SHS ont la chance de pouvoir disposer de la très grande infrastructure nationale TGIR Huma-Num qui met à dispositions une myriade de services et logiciels pour répondre à la variété des besoins des chercheurs (Sharedocs, Huma-Num Box, Nakala, Omeka, Isidore, mais aussi des applications comme ImageMagick, Kakadu, ABBYY, R, Python, Gitlab). Pris indépendamment, chacun de ces services ne peut répondre que partiellement aux principes FAIR. Par exemple, c'est le service Nakala qui permet l'attribution de « handle » pour une identification pérenne de chaque ressource, tandis qu'Isidore augmentera grandement le volet « Findable » du principe FAIR. Ce n'est que par un subtil assemblage de ces différents composants à différents moments de la construction du corpus que nous pouvons nous approcher du paradigme FAIR.

Notre chaîne exploite au maximum les services distants proposés par Huma-Num et notre travail se concentre sur la création de liens indépendants entres services pour s'adapter aux corpus à traiter. Notre proposition implique le développement de « middlewares » dans une architecture logicielle distribuée. Nous développons actuellement des logiciels de liaisons minimalistes (échangeables, facilement réécrits/modifiables et open source) pour transformer et échanger les données entres services Huma-Num. Si un service disparait au profit d'un autre ou n'est tout simplement pas utile pour un corpus spécifique, il suffit soit de récrire le middleware correspondant, soit de l'adapter ou de ne pas l'utiliser. Nous cherchons la souplesse et la maintenabilité en tentant d'éviter le plus possible le développement d'une plateforme « tout-en-un » spécifique.

La chaine automatisée que nous expérimentons a d'ores et déjà permis de traiter des corpus en littérature, archéologie, et bientôt en linguistique. Cette chaine commence à la sortie des scanners de la MSHS jusqu'au dépôt et mise à jour dans Nakalona (Nakala+Omeka), puis moissonnage par Isidore en passant entres autres par des phases de conversion des images TIFF en JPEG (ImageMagick ou XnConvert), d'OCR (ABBYY CLI) distant (envoi/réception par FTP, lancement par SSH), assemblage en Pdf, transcription collaborative (Transcrire) et bientôt des calculs statistiques (R) et des contrôles qualités entre chaque traitement. Chacune de ces étapes produit des fichiers numériques aux formats préconisés par le CINES ainsi que des tableurs de suivis (CSV UTF-8). À chaque étape, les données ont donc un fort potentiel de réutilisabilité à long terme.

Au-delà de l'aspect technique de cette approche, ce projet a permis la formalisation d'un schéma fonctionnel accessible à tous les corps de métiers permettant de faciliter le dialogue entre professionnels et la gestion de projet. C'est la formalisation de ce schéma qui nous a fait apparaître les nombreux points communs entre créateurs de corpus de champs disciplinaires différents.

Le Centre de Données de Géothermie Profonde. Un exemple FAIR

Marc Schaming, Jean Schmittbuhl
Institut de physique du globe de Strasbourg-IPGS (UMR 7516, Université de Strasbourg, CNRS), Strasbourg

Alice Frémand, Marc Grunberg
École et observatoire des sciences de la Terre-EOST (UMS830, Université de Strasbourg, CNRS), France

Nicolas Cuenot
Électricité de Strasbourg-Géothermie, Strasbourg

Le centre de données de géothermie profonde (CDGP, https://cdgp.u-strasbg.fr/) a été mis en place en 2016 par le LabEx G-EAU-THERMIE PROFONDE (http://labex-geothermie.unistra.fr/) afin de préserver, archiver et distribuer les données acquises sur les sites de géothermie en Alsace. Grâce au site pilote de recherche de Soultz-sous-Forêts, plus de trente ans de données ont été collectées dans la région, apportant une importante richesse patrimoniale sur la géothermie.

Dès le départ, il a été décidé de suivre les exigences internationales en terme de gestion de données. Aussi, la recommandation FAIR a été suivie afin de favoriser la découverte, l'accès, l'interopérabilité et la réutilisation des données.

La découverte des données se fait par l'intermédiaire d'une Infrastructure de Données Spatiales (IDS) basée sur GeOrchestra. Les métadonnées sont rassemblées dans un catalogue (GeoNetwork) et suivent les normes INSPIRE/ ISO 19115/19139, spécifiques aux données spatiales. Les données à distribuer sont géo-localisées et correspondent à des expériences effectuées sur les sites de géothermie : il s'agit de données hydrauliques, sismologiques ou géophysiques. Cependant, le besoin de distribuer des simulations numériques a remis en question ce choix, et certains champs ont dû être adaptés. Les données sont identifiées par un identifiant DOI qui est spécifié dans les métadonnées. Les métadonnées sont aussi moissonnées par la plateforme EPOS (TCS-AH https://tcs.ah-epos.eu/).

L'accès aux données est soumis à un protocole AAA permettant l'authentification, l'autorisation et la traçabilité des demandes. En effet, certaines données étant industrielles, cette procédure permet la distribution des données en suivant les règles de propriété intellectuelle, l'affiliation des utilisateurs (académiques, industriels, …) et les règles de distribution définies par les propriétaires.

Pour permettre une bonne interopérabilité et réutilisation des données, les formats acceptés sur la plateforme sont des formats ouverts ou largement utilisées au sein de la communauté. Les données ouvertes sont directement accessibles et des licences Creative Commons (CC- BY ou CC-BY-NC) leur sont attachées.

Malgré une mise en place récente du centre de données, certaines données sont anciennes ; il a fallu traiter des formats et des supports obsolètes afin de les convertir et stocker ces données sur des supports plus modernes. L'identification des propriétaires est parfois difficile, mais nécessaire pour obtenir les règles de diffusion. Par la mise en place de procédures simples, renseignées et décrites dans un plan de gestion de données (Data Management Plan), toutes ces tâches auraient pu être évitées. Les exigences pour une future certification CoreTrustSeal sont également suivies.

SSHADE - L'infrastructure Européenne de bases de données en spectroscopie et spectro-photométrie des solides

Bernard Schmitt, Philippe Bollard, Alexandre Garenne, Damien Albert, Lydie Bonal
IPAG, OSUG-DC / UGA – CNRS, Grenoble

et les partenaires du consortium SSHADE

L'infrastructure Européenne de données en spectroscopie des solides SSHADE (www.sshade.eu), développée dans le cadre du programme Européen Europlanet2020-RI et mise en ligne en début d'année, a pour but de mettre à disposition de la communauté scientifique un ensemble de bases de données « locales » spécialisées sur les propriétés spectroscopiques et photométriques des solides (glaces, minéraux, solides organiques ou inorganiques, matériaux organiques complexes, cosmomatériaux…) et centrées initialement sur des intérêts planétologiques, astrophysiques mais aussi terrestres. Elle devrait aussi intéresser les opticiens, physiciens ou physico-chimistes du solide utilisant la spectroscopie ou nécessitant des données spectro-photométriques.

Les chercheurs impliqués (75 personnes, 21 équipes, 10 pays, dans cette première phase) ont répondu initialement à un appel à intérêt pour mettre en ligne leurs données en leur offrant l'accès à une base de données centralisée à l'OSUG Data Center (Grenoble) mais dont ils conserveraient l'entier contrôle et responsabilité scientifique. Et ceci en leur fournissant tous les outils de préparation, d'import et de validation technique des données ainsi qu'une interface web de recherche, visualisation et export des données par les utilisateurs. Leurs intérêts ont été la forte augmentation attendue de visibilité de leurs données et publications associées, la pérennité des données et surtout métadonnées, une clarification de leur propriété intellectuelle (data référence, DOI, …), et ceci sans aucun développement technique de leur part.

Le problème essentiel dans ce long développement a été de créer et implémenter un modèle de données (SSDM) capable de décrire le plus précisément possible l'ensemble des types de solides couverts par nos domaines variés, ainsi que les différents types de mesures et données spectrales sur l'ensemble du spectre électromagnétique, des rayonnements gamma aux ondes radio. Ceci a été en grande partie obtenu en développant une structure centrale unifiée de description d'un échantillon solide, quel qu'il soit, mais en aménageant des options pour s'adapter au plus près aux domaines scientifiques spécifiques. Il a été fait de même pour les données et produits spectroscopiques et spectro-photométriques, dont la maturation est en cours de finalisation.

La première difficulté résultante de cette exhaustivité est la complexité du modèle de données avec de nombreuses options pour satisfaire aux différents types de données, ce qui nécessite un certain investissement initial des fournisseurs de données, et donc un fort investissement de notre part pour assurer leur formation et leur soutien technique. La seconde a été de trouver un mode de fouille des données simple mais efficace, ce qui semble-t-il a été atteint, mais a nécessité pas mal d'efforts scientifiques et technique. La troisième a été de développer une interface permettant une présentation unifiée mais simple de la structure des échantillons et des données spectrales.

La contrepartie positive de cette structure unique de description (qui pourrait devenir ou inspirer un standard européen ou international dans le domaine) et de stockage des données est la puissance de recherche inter-bases des données sur une multiplicité de critères communs ou spécifiques couvrant tous les domaines. Finalement, un développement méthodologique et technique unique, mais en concertation, pour une large communauté est une solution efficace, mais nécessite en contrepartie un soutien spécifique.

dat@OSU, une plateforme de référencement et de valorisation des données de la recherche à l'OSU THETA

Hélène Tisserand
OSU THETA de Franche-Comté/Bourgogne (UMS3245) - Observatoire de Besançon

Sylvie Damy
Laboratoire Chrono-environnement (UMR6249)

Bernard Debray
Institut UTINAM (UMR6313)

Raphaël Melior
OSU THETA de Franche-Comté Bourgogne (UMS3245)

L'expérience présentée ici concerne le projet de gestion et valorisation des données de la recherche de l'Observatoire des Sciences de l'Univers THETA de Franche-Comté Bourgogne. Ce projet, initié en 2013 pour un démarrage effectif en 2014, avait pour objectif de développer des pratiques liées au mouvement de l'open data. Il a abouti à la mise en place d'un portail de métadonnées, dat@OSU, pour décrire et rendre visibles les données générées dans les laboratoires et équipes de l'OSU THETA.

L'OSU THETA regroupe des thématiques/disciplines très diverses (astronomie, sciences de la Terre, sciences de l'environnement, physique, archéologie, chimie, …). Dans ce contexte multidisciplinaire, il n'était pas possible de gérer le stockage et l'organisation des données qui restent du ressort des laboratoires. L'OSU THETA, en tant que structure fédérative, s'est donc focalisé sur la définition d'une description standardisée de ces données, en mutualisant les expériences, les moyens et les savoir-faire de chaque discipline. Le portail dat@OSU, qui en résulte, met à disposition sur le web les métadonnées (identifiées de manière pérenne - DOI) et est interopérable avec d'autres portails (Isidore, DataCite, …).

Après un état de l'art initial sur l'open research data, un profil de métadonnées a été conçu ; il utilise des standards généralistes (Dublin Core, DataCite, …) et disciplinaires et permet ainsi l'interopérabilité. Afin de décrire les données, des thésaurus adoptés par les différentes communautés scientifiques ont été utilisés. Le portail dat@OSU a été développé à partir de ce profil et inauguré en avril 2016. Il propose actuellement plusieurs centaines de fiches de métadonnées. Son développement a été un point d'appui pour l'organisation d'un colloque sur les données de la recherche sous l'égide de la COMUE Université Bourgogne Franche-Comté (DataBFC, novembre 2017). La seconde édition de ce colloque est prévue pour 2019.

Le projet s'est appuyé sur une forte collaboration inter-métiers : Chercheur, Informaticien, Documentaliste.

Des chercheurs ont participé dès les premières étapes en faisant remonter leurs besoins ; des référents ont, en particulier, été impliqués pour chaque laboratoire.

L'OSU a largement mis à disposition son ingénieur informaticien pour le développement. Le projet a de plus bénéficié du soutien de l'OSU, des laboratoires, des tutelles (Université, CNRS) et de la région, en particulier pour l'embauche en CDD d'une documentaliste. L'INSU a attribué en juillet 2017 un poste de documentaliste pour le projet.

L'organisation de «metadata parties» ou de rendez-vous individuels a permis de dépasser les réserves de certains chercheurs et de sensibiliser la communauté aux bonnes pratiques en matière de gestion des données de la recherche.

La taille de la structure porteuse (environ 600 personnes) et le soutien fort des directions et des tutelles ont permis de réaliser le projet dans des conditions très favorables. Des contacts avec d'autres structures de recherche (OSU, COMUE UBFC) sont en cours et devraient aboutir au passage à l'échelle et à la diffusion du projet.

Contributeurs

(par ordre alphabétique)

Albert	Damien	damien.albert@univ-grenoble-alpes.fr
André	Francis	francis.andre@cnrs-dir.fr
Beckmann	Volcker	vbeckmann@admin.in2p3.fr
Bollard	Philippe	philippe.bollard@univ-grenoble-alpes.fr
Bonal	Lydie	lydie.bonal@univ-grenoble-alpes.fr
Bonamy	Cyrille	cyrille.bonamy@univ-grenoble-alpes.fr
Bourgès	Laurent	laurent.bourges@univ-grenoble-alpes.fr
Braud	Isabelle	isabelle.braud@irstea.fr
Brunel	Valentin	valentin.brunel@sciencespo.fr
Bryas	Emmanuelle	emmanuelle.bryas@inrap.fr
Cadorel	Sarah	sarah.cadorel@inalco.fr
Chaffard	Véronique	veronique.chaffard@univ-grenoble-alpes.fr
Chauchat	Julien	julien.chauchat@univ-grenoble-alpes.fr
Chesnet	David	david.chesnet@mshs.univ-poitiers.fr
Chester	Chloë	chloe.chester@mnhn.fr
Colin	Camille	camille.colin@inrap.fr
Côtez	Emmanuel	emmanuelle .cotez@mnhn.fr
Coussot	Charly	charly.coussot@univ-grenoble-alpes.fr
Cuenot	Nicolas	nicolas.cuenot@es.fr
Damy	Sylvie	sylvie.damy@univ-fcomte.fr
Dassas	Karin	karin.dassas@ias.u-psud.fr
Debray	Bernard	bernard.debray@utinam.cnrs.fr
Etienne	Carole	carole.etienne@ens-lyon.fr
Fabre	Juliette	juliette.fabre@umontpellier.fr
Frémand	Alice	almand@bas.ac.uk
Fromont	Emilie	emilie.fromont@sciencespo.fr
Fusberti	Benjamin	benjamin.fusberti@ganil.fr
Galle	Sylvie	sylvie.galle@ird.fr
Garenne	Alexandre	alexandre.garenne2@gmail.com
Glorian	Jean-Michel	jean-michel.glorian@irap.omp.eu
Gomez-Diaz	Teresa	teresa.gomez-diaz@univ-mlv.fr
Grunberg	Marc	marc.grunberg@unistra.fr

Contributeurs

(par ordre alphabétique)

Heintz	Wilfried	wilfried.heintz@inra.fr
Jacob	Daniel	daniel.jacob@inra.fr
Jomier	Rémy	remy.jomier@mnhn.fr
Juen	Patrick	patrick.juen@univ-grenoble-alpes.fr
Le Bras	Yvan	yvan.le-bras@mnhn.fr
Le Sidaner	Pierre	pierre.lesidaner@obspm.fr
Lesteven	Soizick	soizick.lesteven@astro.unistra.fr
Liegois	Loic	loic.liegeois@univ-paris-diderot.fr
Lobry	Olivier	olivier.lobry@umontpellier.fr
Mabille	Anne	anne.mabille@mnhn.fr
Mathieu	Antoine	antoine.mathieu@univ-grenoble-alpes.fr
Melior	Raphaël	raphael.melior@obs-besancon.fr
Meunier	Jean-Charles	jean-charles.meunier@lam.fr
Moreau	Anne	anne.moreau@inrap.fr
Moreau	Gabriel	gabriel.moreau@legi.cnrs.fr
Moulard	Nathalie	nathalie.moulard@cnrs-orleans.fr
Nauge	Michael	michael.nauge@univ-poitiers.fr
Nouvel	Blandine	blandine.nouvel@frantiq.fr
Parisse	Christophe	cparisse@parisnanterre.fr
Poulain	Pierre	pierre.poulain@univ-paris-diderot.fr
Quidoz	Marie-Claude	marie-claude.quidoz@cefe.cnrs.fr
Rivet	Alain	alain.rivet@cermav.cnrs.fr
Romier	Geneviève	genevieve.romier@cc.in2p3.fr
Rousset	Miled	miled.rousset@mom.fr
Sanguillon	Michèle	michele.sanguillon@univ-montp2.fr
Schaaff	Andé	andre.schaaff@astro.unistra.fr
Schaming	Marc	marc.schaming@unistra.fr
Schmitt	Bernard	bernard.schmitt@univ-grenoble-alpes.fr
Schmittbuhl	Jean	jean.schmittbuhl@unistra.fr
Sommeria	Joël	joel.sommeria@univ-grenoble-alpes.fr
Tisserand	Hélène	helene.tisserand@univ-fcomte.fr
Tuffery	Christophe	christophe.tuffery@inrap.fr

Mission pour les Initiatives Transverses et Interdisciplinaires
Plateforme réseaux - L'inter-réseaux

Groupe de travail inter-réseaux - Atelier Données

Ce groupe de travail, composé de représentants de plusieurs réseaux, s'attache à construire la base d'une réflexion sur la gestion des données porteuse d'une vision « métiers » et « réseaux ». À partir d'une réflexion de type « cycle de vie de la donnée », il se propose de cartographier les usages dans chaque réseau autour de la gestion de la donnée afin de :

- Construire et diffuser une vision transversale de la gestion des données pour enrichir la pratique de chaque réseau dans le domaine des données et permettre le développement de la complémentarité entre réseaux ;
- Valoriser l'apport des expériences et expertises « métier » dans une vision transversale de gestion de données dans les réseaux technologiques et scientifiques de la MITI ;
- Sensibiliser les communautés professionnelles de l'appui à la gestion des données (organisation de journées thématiques par exemple) ;
- Identifier les problématiques concernant les « data » dans chaque réseau (livrables à définir) ;
- Mettre en commun et partager de nouvelles pratiques en réseau et au sein de chaque réseau.

pour plus d'informations
https://www.cnrs.fr/mi/spip.php?article1313

contact GT inter-réseaux Ateliers Données
gt-donnees-inter-reseaux@groupes.renater.fr

Direction Information scientifique et technique