Résumé des interventions

Lundi 13 novembre 2017

Francis André (CNRS - Direction de l'information scientifique et technique (DIST)) **:** « Gestion des données de recherche au CNRS : stratégie institutionnelle et approches disciplinaires »
Rendre FAIR (Findable, Accessible, Interoperable, Reusable) les données de recherche est une action essentielle du processus de la science ouverte.
Atteindre cet objectif nécessite la mobilisation de ressources et de compétences multiples, dont l'efficience sera d'autant plus grande que la stratégie qui la porte est transparente et affirmée.
La présentation se propose d'analyser les divers paramètres à prendre en compte dans une démarche volontariste vers de bonnes pratiques de gestion de données (technique, politique, organisationnel, juridique, éthique,…) et d'en mesurer l'impact dans la définition de stratégies institutionnelles et disciplinaires.

Mohamed Yahia (CNRS – INIST) **:** « Datacite : Infrastructure pour valoriser et rendre FAIR les données de la recherche »
La valorisation et le partage des données de la recherche est une composante principale du mouvement de l'open access et de l'open science. Eléments importants de ce mouvement, les identifiants pérennes facilitent la découverte et l'accès sur le long terme aux produits de la recherche scientifique : Ils permettent d'identifier, de référencer et de citer ces produits. Ils concourent ainsi à leur visibilité, à leur partage et à les rendre FAIR.
Le consortium DataCite a mis en place une infrastructure, basée sur le DOI, spécifiquement dédiée aux données de la recherche pour favoriser leur découverte, leur accessibilité et leur citation. Cette intervention sera consacrée à la présentation des services développés par le consortium DataCite autour du DOI ; et à la procédure à suivre pour attribuer des DOI à vos jeux de données.

Eric Bouillet (Swiss Data Science Center) : « Ouverture et gestion des données de la recherche vus et entrepris par Swiss Data Science Center »
Dans le cadre de cette présentation, Eric Bouillet du Swiss Data Science Center (SDSC) parlera des atouts et des défis associés à la digitalisation dans le monde de la recherche. La promotion de la collaboration multidisciplinaire, l'ouverture contrôlée de la donnée, les besoins de transparence et de reproductibilité sous respect des directives européennes seront parmi les thèmes abordés. Pour conclure, Eric parlera des méthodes développées par le SDSC pour mener à bien ces défis.

Christophe Cruz (Laboratoire Electronique, Informatique et Image - UMR 6306) **:** « Problématique Big Data : définition et enjeux »
La question du Big Data est aujourd'hui centrale dans de nombreux domaines d'activités industrielles et académiques. Après une évolution rapide des technologies numériques, le périmètre et la définition du Big Data se font plus difficiles à établir tant la variété des applications et des outils sont pléthores. Le Big Data ne se limite pas aux données massives, mais se caractérise par un changement de paradigme prenant part à la révolution numérique. Ainsi les enjeux sont majeurs par ses défis et opportunités pour les métiers de l'enseignement et la recherche. Cette intervention se propose de présenter quelques facettes du Big Data afin d'apporter des définitions contextualisées ainsi que quelques enjeux pour l'enseignement supérieur et de la recherche.

Mardi 14 novembre 2017

Bernard Sampité (CNRS - INIST) **:** Le portail OPIDoR : des outils et services pour optimiser le partage et l'interopérabilité des données de la recherche »
L'Inist-Cnrs a mis en place, à destination de l'ESR, un ensemble d'outils et de services pour optimiser le partage et l'interopérabilité des données de la recherche. DMP-OPIDoR, Cat-OPIDoR et le service Datacite vous permettent de gérer vos données dans le cadre de votre projet de recherche.
Présentation des enjeux d'une bonne gestion des données de la recherche et des services du portail OPIDoR.

Lorène Béchard (CINES) **:** « Gestion et préservation des données de la recherche : l'offre de service du CINES »
De nos jours la production d'informations est essentiellement numérique et les exigences autour de la gestion des données sont de plus en plus prégnantes : big data, propriété, interopérabilité, open data, réutilisation, etc. Pour y répondre efficacement, cela nécessite de s'interroger sur les moyens de pérennisation de ces données.
L'archivage numérique pérenne consiste à conserver le document et l'information qu'il contient, dans son aspect physique comme dans son aspect intellectuel, de manière à ce qu'il soit accessible et compréhensible aussi longtemps que nécessaire.
Mandaté par son ministère de tutelle, le CINES est depuis 2004 l'opérateur pour l'archivage des données et documents numériques produits par la communauté Enseignement supérieur et Recherche française. Il propose des solutions d'archivage numérique sur le moyen et long terme, mutualisées, performantes, certifiées, économiques et personnalisables.

Stéphane Pouyllau (CNRS) **:** « Huma-Num et son offre de service pour les SHS »
Huma-Num est une très grande infrastructure de recherche (TGIR) visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales.
Pour remplir cette mission, la TGIR Huma-Num est bâtie sur une organisation originale consistant à mettre en œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs.
La TGIR Huma-Num favorise ainsi, par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques, la coordination de la production raisonnée et collective de corpus de sources (recommandations scientifiques, bonnes pratiques technologiques).
Elle développe également un dispositif technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche. Ouvert à l'ensemble des programmes de recherche de l'enseignement supérieur et de la recherche (UMR, UMS, EA, etc.), il est composé de services numériques dédiés, d'une plateforme d'accès unifié (ISIDORE) et d'une procédure d'archivage à long terme.
La TGIR Huma-Num propose en outre des guides de bonnes pratiques technologiques généralistes à destination des chercheurs. Elle peut mener ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans les ERIC (European Research Infrastructure Consortium) DARIAH et CLARIN en coordonnant les contributions nationales. Elle est également impliquée depuis 2015 dans deux projets H2020 : Parthenos et Humanities at Scale.

Chloé Martin (BBEES - UMS 3468) **:** « UMS BBEES – Son rôle dans la gestion et l'ouverture des données de la recherche »
Créée fin 2011, l'unité mixte de service CNRS-INEE et MNHN a pour objectif d'aider les chercheurs à organiser et administrer leurs données tout en facilitant leur mise à disposition et leur accessibilité.
Pour cela, l'unité met en place des outils informatiques (bases de données, interface de gestion et de saisies) et propose, au travers d'un portail en ligne, les métadonnées des jeux de données et bases de données existant dans le domaine de l'écologie et de l'environnement. Elle est également à disposition des personnels souhaitant être informés sur l'organisation, la gestion (DMP, DOI, métadonnées, référentiel) et les questions juridiques concernant les données.

Sylvie Damy[(1)], Bernard Debray[(2)], Raphaël Melior[(3)], Hélène Tisserand[(3)], Gaëlle Laporte[(3)] ((1) Chrono-environnement - UMR 6249, (2) UTINAM - UMR 6213, (3) OSU-THETA - UMS 3245) **:** « Dat@OSU : le portail de référencement des données de la recherche de l'OSU THETA »
Le projet Dat@OSU (https://dataosu.obs-besancon.fr/) développé dans le cadre de l'Observatoire des Sciences de l'Univers de Franche-Comté Bourgogne (OSU THETA) vise à permettre la description standardisée des ensembles et bases de données scientifiques produites par les laboratoires et équipes de recherche associées à l'OSU. Par le biais d'une fiche de métadonnées, construites sur un modèle compatible avec les standards de métadonnées généralistes (Dublin Core, DataCite) et disciplinaires (observatoire virtuel astronomique, GBIF pour les données environnementales, etc) le portail Dat@OSU permet de décrire les données dans 4 grands groupes d'information : générales (disciplines ; mots-clés ; couvertures spatiale, temporelle, spectrale, ...), administratives (créateurs, contributeurs, projets associés, ...), technique (méthode d'obtention, formats, ...) et structurelle (groupement en collections, ...).

Table ronde :

Benjamin Pohl (Biogéosciences - UMR 6282) **:** « Acquisition – Valorisation – Archivage – Partage de la Donnée Scientifique au Laboratoire Biogéosciences »
Le laboratoire Biogéosciences (CNRS / univ. Bourgogne / EPHE) est une unité mixte de recherche pluri-disciplinaire qui produit de grandes quantités de données scientifiques hétérogènes dans leur format, leur volumétrie, leur utilisation et leur finalité. Il s'agit ici d'illustrer quelques actions entreprises par le laboratoire pour garantir leur archivage sécurisé sur le long terme, leur valorisation et leur affichage auprès de la communauté scientifique ou extra-académique, leur traitement voire leur partage (public ou avec des utilisateurs identifiés).

Vincent Boudon (Laboratoire Interdisciplinaire Carnot de Bourgogne - UMR 6303) **:** « Les données en spectroscopie moléculaire et le projet VAMDC »
Il existe un grand nombre de bases des données en spectroscopie moléculaire. Celles-ci sont essentielles pour les applications atmosphériques et planétologiques, mais il est parfois difficile de s'y retrouver pour les utilisateurs. Le projet VAMDC (Virtual Atomic and Molecular Data Centre) vise à répondre à cette problématique en permettant l'accès simultané à toutes les sources de données existantes pour une molécule donnée. Dans ce cadre, l'équipe SMPCA à Dijon développe ses propres bases, mais aussi des outils pour le portail (http://portal.vamdc.org) pour l'export des données dans différents formats et pour leur visualisation. Les bases développées sont également référencées par le projet Dat@OSU.

Francis Raoul, Sylvie Damy, Jean-Daniel Tissot, Charles-Henri Falconnet (Chrono-environnement - UMR 6249) **:** « Gestion des données au laboratoire Chrono-environnement: enjeux et pratiques »
Depuis 2012, le laboratoire Chrono-environnement a mis en place un axe transversal "Bases de données" qui vise à accompagner les projets de gestion de jeux et bases de données de recherche, issues de multiples sources (observations de terrain, de laboratoire, simulations…). Ces projets sont développés, en partenariat étroit avec l'OSU THETA, pour répondre au mieux aux attentes des collègues, allant de la simple description des métadonnées sur la plateforme dat@osu à la création d'une application ouverte en ligne. Cette présentation a pour but de partager les enjeux et pratiques autour des données de la recherche dans ce laboratoire pluridisciplinaire.

Marc Steinmann (Chrono-environnement - UMR 6249) **:** « Les besoins en gestion des données dans la Zone Atelier Arc Jurassien (ZAAJ) »
La ZAAJ coordonne des recherches interdisciplinaires à long terme sur l'environnement et les écosystèmes jurassiens, en relation avec les questions sociétales dans l'arc jurassien. Les données produites par les différentes thématiques sont variées avec des besoins très spécifiques en terme de gestion des données. Une synthèse de ces besoins sera présentée sur la base des fiches de métadonnées du projet Dat@OSU, complétée par des exemples plus précis.

Philippe Rousselot (UTINAM - UMR 6213) : « L'OSU THETA : intérêt d'une structure fédérative pour développer une bonne gestion des données de recherche »
Le rôle d'un Observatoire des Sciences de l'Univers sur la question de la gestion des données sera présenté ; on expliquera ce qu'une telle structure peut apporter à ce type de problématique, en évoquant notamment le projet Dat@OSU présenté par ailleurs.

Retours d'expérience

Ernest Chiarello (ThéMA - UMR 6049), **Marion Landré** (MSHE Ledoux - USR 3124), Damien Roy (ThéMA - UMR 6049) **:** « Retour d'expérience sur la gestion des données à référence spatiales et de leurs métadonnées - utilisation d'OwnCloud pour publier des données cartographiques et leurs métadonnées dans geOrchestra »
Les données spatiales produites par les laboratoires fédérés à la MSHE manquent de visibilité et sont peu partagées. Afin de renforcer leur centralisation et le catalogage des métadonnées qui leur sont associées, une Infrastructure de Données Spatiales (IDS) basée sur geOrchestra a été mise en place à la MSHE.
geOrchestra intègre plusieurs composants cartographiques dans une IDS libre, riche en fonctionnalités, avec en particulier une application dédiée à la publication des données cartographiques, GeoServer, et un gestionnaire des métadonnées, GeoNetwork. En parallèle, il a été choisi de déployer un serveur de stockage et de partage de fichiers en ligne, Owncloud, l'objectif étant de relier les deux outils, geOrchestra et OwnCloud, de manière à simplifier la chaîne opératoire allant de la production à la valorisation des données à références spatiales.
Le projet ge@sync vise ainsi la publication automatique en ligne des données cartographiques dès qu'elles sont déposées et partagées sur le serveur OwnCloud. Les données et leurs métadonnées sont gérées uniquement sur les postes des utilisateurs avec des applications métier telles que QGIS ou ArcGIS. Par construction, il n'y a pas de processus de validation de la qualité des métadonnées, elles sont envoyées telles que sur le serveur de publication, libre alors à leurs propriétaires de les corriger si elles sont incomplètes ou comportent des erreurs.
Les difficultés proviennent du fait que données et métadonnées sont envoyées sur des applications différentes, GeoServer et GeoNetwork, ce qui induit des écarts puisqu'on peut se retrouver avec des données sans leurs métadonnées et inversement.
Ce travail a été réalisé par l'équipe Géomatique de la MSHE en collaboration avec le laboratoire ThéMA.

José Lages (UTINAM - UMR 6213) **:** « Analyse de réseaux issus de (méga) données »
La plupart des données issues des activités humaines peuvent être vues comme des réseaux complexes dirigés. Nous présenterons des modèles théoriques, notamment la méthode de la matrice de Google, permettant d'extraire de l'information pertinente et non triviale de ces grandes masses de données.

Hervé Richard (Chrono-environnement - UMR 6249) **:** « PollenChrono : base de données des analyses polliniques »
Les sédiments lacustres, palustres et tourbeux, et certains sédiments déposés dans les sites archéologiques, conservent un nombre plus ou moins importants de grains de pollen et de spores. Leur extraction des sédiments, leur détermination et leur comptage (ensemble de processus appelé ˝analyse pollinique˝ ou ˝palynologie˝) permettent de proposer une reconstitution de la végétation à une époque donnée. Les différentes phases reconnues ainsi dans l'évolution de la végétation au cours du temps seront dépendantes de la variation des paramètres climatiques et de l'impact de l'homme.
La base de données PollenChrono regroupe l'ensemble des analyses polliniques effectuées depuis 1981 par le laboratoire de Besançon (aujourd'hui "Laboratoire Chrono-environnement"). Une fiche est élaborée pour chaque site (un site correspondant à un point d'analyse). Cette fiche indique le nom de la commune et du lieu-dit, les coordonnées du site (latitude, longitude, altitude), son emplacement exact sur une carte, les informations principales concernant l'analyse (nombre d'échantillons, qualité des résultats, résultats marquants…), la ou les périodes chronologiques représentées, le nom et l'affiliation du ou des créateur(s) (personne(s) ayant analysé les échantillons), le nom et l'affiliation du ou des contributeur(s) (personne(s)

travaillant sur le site ou ayant été impliquée(s) dans la collecte des échantillons), les principales références (articles, ouvrages, rapports) où ces résultats ont été publiés, consignés ou exploités. Dans un second temps, un ou des fichiers pdf seront associés à ces fiches. Ils permettront l'accès direct au comptage pour chaque échantillon (feuilles de comptage en données brutes), fichiers regroupant l'ensemble des échantillons d'un site, diagrammes, publications… À ce jour, environ 250 fiches ont été validées par l'équipe Dat@OSU ; elles sont classées par ordre alphabétique des communes.
https://dataosu.obs-besancon.fr/edit_ref.php?ref=collection&id=13

Jérôme Thomas (Biogéosciences - UMR 6282) : « E-ReColNat : plateforme informatique du Réseau des Collections Naturalistes »
Le programme e-ReColNat est coordonné par un consortium regroupant le Muséum National d'Histoire Naturelle de Paris (MNHN), les universités de Bourgogne, de Clermont-Ferrand et de Montpellier, l'Institut National de Recherche en Agronomie (INRA), l'Institut de Recherche pour le Développement (IRD), le Conservatoire National des Arts et Métiers (CNAM), l'association Tela Botanica, la société Agoralogie, le point nodal GBIF-France (Global Biodiversity Information Facility) et le Centre National pour la Recherche Scientifique (CNRS). Il est lauréat des Programmes d'Investissements d'Avenir (PIA) pour les Infrastructures en biotechnologie et santé.
Le nombre de spécimens conservés dans les différentes collections naturalistes françaises (muséums, musées et universités) est estimé à plus de 100 millions. Il était important de mettre en place un Réseau des Collections Naturalistes (ReColNat) afin de coordonner les inventaires et offrir une réponse globale pour la valorisation internationale de ces collections. Ce réseau est labellisé Infrastructure de Recherche (IR) par le Ministère de l'Enseignement Supérieur et de la Recherche.
Ce programme hérite des expériences acquises antérieurement (par exemple la documentation participative des spécimens botaniques, l'informatisation et la numérisation des référentiels paléontologiques). E-ReColNat est la couche informatique de ReColNat et a pour vocation d'extraire et mettre à disposition ces données naturalistes par 1) une numérisation massive et une documentation par la science participative, 2) l'informatisation et la numérisation des spécimens de référence en paléontologie et zoologie et 3) le moissonnage des données.
Le laboratoire Biogéosciences de l'Université de Bourgogne coordonne nationalement la vérification et l'informatisation des données, la photographie et le traitement des images des spécimens types et figurés paléontologiques et zoologiques repérés dans les collections. Le récent développement d'une interface de saisie permet à tout établissement détenteur de verser massivement ou spécimen par spécimen des données d'inventaire. Le versement massif de données permet ainsi de valoriser l'intégralité de l'inventaire de collections paléontologiques et zoologiques et non uniquement les types et figurés. Ce système est également ouvert aux particuliers et aux associations souhaitant diffuser le contenu de leurs collections. Ces informations sont accessibles sur un seul et même portail national (www.recolnat.org). Ce système permet de centraliser les données issues des différents partenaires en utilisant des standards internationaux de données naturalistes (Darwin Core) afin de favoriser la qualité, le partage et la stabilité des données. Le portail héberge des outils consultatifs ainsi que des outils participatifs. Ces applications, comme les Herbonautes, permettront aux experts naturalistes de compléter les inventaires et de documenter les spécimens. Ces nouvelles données pourront être moissonnées afin de compléter les inventaires. Actuellement axé sur les collections paléontologiques, botaniques et zoologiques, ReColNat devrait prochainement s'étendre aux autres domaines des sciences naturelles

Mercredi 15 novembre 2017

Françoise Genova (Observatoire Astronomique de Strasbourg - UMR 7550) **:** « Les données astronomiques et l'Observatoire Virtuel »
L'astronomie a été à l'avant-garde du partage des données scientifiques, qui constituent l'une des infrastructures de recherche de la discipline. Les activités du Centre de Données astronomiques de Strasbourg, dont les services sont largement utilisés par la communauté scientifique internationale, sont brièvement décrites, ainsi que la mise en réseau des services en ligne, puis le développement du cadre d'interopérabilité disciplinaire, l'Observatoire Virtuel (OV) astronomique, dans l'International Virtual Observatory Alliance. Les aspects internationaux, européens et internationaux de l'OV sont évoqués. Grâce aux efforts des producteurs de données et des développeurs de l'OV, les données astronomiques sont FAIR.

Hélène Skrzypniak : "Les données de la recherche : droits et devoirs du chercheur "
Ces dernières années, les textes destinés à promouvoir l'ouverture des données de la recherche se sont multipliés, tant au niveau international que national. Le mouvement est aujourd'hui désigné sous les termes d'open access. Il vise, notamment, à encourager la publication en libre accès des publications scientifiques. La France n'a pas été épargnée par ce mouvement comme en atteste l'adoption récente de la loi pour une République numérique du 7 octobre 2016 dont plusieurs dispositions visent à favoriser l'ouverture des données de la recherche.

Ces différents textes créent ainsi des nouvelles obligations pour le chercheur, mais aussi de nouveaux droits. Dans ce contexte, nous proposons de revenir sur le cadre juridique de l'ouverture des données de la recherche : quelles sont les données concernées ? Quelles conditions doivent être respectées ? L'ouverture des données est-elle une obligation pour le chercheur ? Quelle protection pour le chercheur qui ouvre ses données ?

Françoise Genova (Observatoire Astronomique de Strasbourg - UMR 7550) **:** « La Research Data Alliance »
La Research Data Alliance (RDA, https://www.rd-alliance.org/) est une organisation internationale qui a pour objectif de faciliter le partage des données de la recherche. Créée en 2013, elle a actuellement plus de 6200 membres de profils très divers, qui mènent des activités très variées qui seront brièvement présentées.