



HAL
open science

Guide de bonnes pratiques sur la gestion des données de la Recherche

Christine Hadrossek, Joanna Janik, Maurice Libes, Violaine Louvet,
Marie-Claude Quidoz, Alain Rivet, Geneviève Romier

► To cite this version:

Christine Hadrossek, Joanna Janik, Maurice Libes, Violaine Louvet, Marie-Claude Quidoz, et al..
Guide de bonnes pratiques sur la gestion des données de la Recherche. 2021. hal-03152732v1

HAL Id: hal-03152732

<https://hal.science/hal-03152732v1>

Submitted on 25 Feb 2021 (v1), last revised 1 Feb 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Guide de Bonnes Pratiques sur la gestion des données de la Recherche

Groupe de travail "Atelier Données"

Mission pour les Initiatives Transverses Interdisciplinaires

févr. 09, 2021

Ce guide de bonnes pratiques sur la gestion des données dans les réseaux métiers, a été réalisé par :

- Christine Hadrossek : DDOR, réseau Renatis.
- Joanna Janik : DDOR, réseau Renatis.
- Maurice Libes : réseau SIST.
- Violaine Louvet : réseau Calcul.
- Marie-Claude Quidoz : réseau rBDD.
- Alain Rivet : réseau QeR.
- Geneviève Romier : réseau rBDD.

Pour contacter les auteurs, envoyez un message à contact-guide@services.cnrs.fr.

Mise en page et logo - Pierre Navaro

Cet ouvrage est mis à disposition selon les termes de la licence Creative Commons CC BY 4.0 

<https://mi-gt-donnees.pages.math.unistra.fr/guide>

Table des matières

| | | |
|----------|------------------------------------------------------------------------------------------------------------------------------|-----------|
| 1 | Imaginer et préparer | 5 |
| 1.1 | Comprendre le paysage de la gestion des données | 5 |
| 1.2 | Comprendre et respecter la législation en vigueur | 9 |
| 1.3 | Connaitre et comprendre les principes FAIR | 10 |
| 1.4 | Prévoir la traçabilité des données | 11 |
| 1.5 | Envisager la curation des données | 12 |
| 1.6 | Prévoir l'archivage des données | 13 |
| 1.7 | Identifier les compétences et expertises pour la gestion des données de la recherche | 13 |
| 2 | Concevoir et planifier | 15 |
| 2.1 | Évaluer les besoins liés au projet | 16 |
| 2.2 | Mettre en place une gestion de projet | 16 |
| 2.3 | Amorcer un plan de gestion de données | 19 |
| 2.4 | Identifier les infrastructures adaptées au projet : fournisseur du service, fonctionnalités, capacités et services | 21 |
| 3 | Collecter | 23 |
| 3.1 | Utiliser des normes et des standards d'interopérabilité | 24 |
| 3.2 | Les systèmes d'acquisition : maîtriser l'acquisition et la collecte des données | 27 |
| 3.3 | Environnements de stockage - Sauvegarder les données | 33 |
| 4 | Traiter | 35 |
| 4.1 | Préparer les fichiers de données en vue de leur analyse | 35 |
| 4.2 | Organiser les données | 37 |
| 4.3 | Mettre en place un contrôle qualité des données | 40 |
| 5 | Analyser | 45 |
| 5.1 | Plateformes de traitement de données | 45 |
| 5.2 | Outils pour l'analyse des données | 46 |
| 5.3 | Mettre en place des méthodes d'analyse et des chaînes logicielles | 48 |
| 5.4 | Optimiser l'utilisation des ressources | 51 |
| 5.5 | Traitements sémantiques/ linguistiques | 52 |
| 6 | Préserver et archiver | 55 |
| 6.1 | Comprendre et différencier les différents concepts | 55 |
| 6.2 | Préserver les objets numériques | 58 |
| 6.3 | Archiver les objets numériques | 60 |

| | | |
|-----------|----------------------------------------------------------------------------------------|-----------|
| 6.4 | Sélectionner les données pertinentes | 61 |
| 6.5 | S'appuyer sur les enseignements des retours d'expérience | 62 |
| 7 | Publier et diffuser | 65 |
| 7.1 | Communiquer et documenter | 65 |
| 7.2 | Publier les métadonnées | 66 |
| 7.3 | Utilisation de thesaurus | 69 |
| 7.4 | Utilisation d'identifiants pérennes | 70 |
| 7.5 | Les entrepôts de données | 73 |
| 7.6 | Publier un « Datapaper » pour valoriser et expliciter les données | 75 |
| 7.7 | Publier des données grâce au web des données et au web sémantique | 78 |
| 8 | Conclusion | 79 |
| 9 | Glossaire | 81 |
| 10 | Infrastructures | 83 |
| 10.1 | Infrastructures Européennes | 83 |
| 10.2 | Infrastructures Nationales | 85 |
| 11 | Reproductibilité | 93 |
| 11.1 | Comprendre les enjeux et défis | 93 |
| 11.2 | Utiliser des environnements et des outils qui favorisent la reproductibilité | 94 |
| 11.3 | Développement open source et reproductibilité | 95 |
| 12 | Autres guides de bonnes pratiques | 97 |
| 13 | Crédits | 99 |
| 13.1 | Auteurs | 99 |
| 13.2 | Relecteurs | 99 |
| 13.3 | Licence | 100 |

De l'importance des données de la recherche

La gestion rigoureuse et cohérente des données de la recherche constitue aujourd'hui un enjeu de taille pour la production de nouvelles connaissances scientifiques. Guidés par le « [Plan National pour la Science Ouverte](#) », les différents organismes de recherche et les Instituts du CNRS s'emparent de ces questions primordiales pour participer à la réflexion et à la mise à dispositions des outils, méthodes et infrastructures répondant aux besoins des communautés scientifiques en matière de gestion et de partage des données scientifiques.

Améliorer les pratiques de gestion des données de la science devient nécessaire pour garantir l'intégrité scientifique et la traçabilité de la recherche produite, mais aussi pour rendre accessible, partager, permettre la réutilisation ou la reproductibilité des données, qui on peut le rappeler sont financées à plus de 50% sur des fonds publics.

Gérer les données de la recherche est un processus complexe qui suppose un travail long et coûteux, des moyens techniques et humains parfois importants et qui comprend plusieurs étapes avant d'aboutir à la publication et l'archivage de données fiables, de qualité, respectueuses du droit des personnes et de la législation en vigueur.



Fig. 1 – Cartographie des actions des réseaux métiers autour de la gestion des données.

Pour formaliser les différentes étapes de gestion des données, nous nous sommes servis du « cycle de vie des données » élaboré au sein de l'Atelier Données. Il s'agit d'un cercle vertueux que l'on peut faire correspondre aux différentes phases d'un projet scientifique.

L'apport des réseaux métiers du CNRS

Dans leurs différentes pratiques, les réseaux métiers du CNRS, regroupés au sein de la [Mission pour les Initiatives Transverses Interdisciplinaires \(MITI\)](#) ou soutenus par les Instituts sont en première ligne pour participer à ce mouvement d'ouverture et de partage des données. Les personnels des organismes de recherche qui les constituent, œuvrent pour mettre en place de bonnes pratiques de gestion et participent également au processus de production des données scientifiques aux côtés des équipes de recherche. C'est aussi dans ce cadre qu'ils interviennent en appui et soutien à la recherche scientifique.

C'est précisément ce travail de soutien que nous nous proposons de rendre compte dans ce document qui, à travers de nombreux séminaires communications et formations, vise à fournir les meilleures pratiques du moment en matière de gestion des données, et peut ainsi s'apparenter à un « guide de bonnes pratiques ».

Certes de nombreux guides existent déjà dans le domaine, mais l'originalité de ce document réside dans son application aux données de la recherche sous l'angle de différents métiers de la recherche. Il fournit donc un point de vue

transversal intéressant et traduit les efforts et le soutien mis en place par les personnels d'appui à la recherche au sein des réseaux, dans la gestion et la valorisation des données scientifiques.

Objectifs du guide

Ce guide est la production du groupe de travail inter-réseaux « *Atelier Données* ». Il s'agit d'un groupe composé de plusieurs réseaux métiers de la MITI (Calcul, Devlog, QeR, rBDD, Renatis, Resinfo, Medici), du réseau SIST, (labellisé par l'INSU et regroupant les gestionnaires de données environnementales), de l'INIST, et de la Direction des données ouvertes de la recherche (DDOR-CNRS¹). Les activités de cet atelier ont été présentées lors du séminaire SIST 2019.

Ce guide fait suite à un premier travail très synthétique réalisé en 2017 qui visait à établir une *cartographie de l'action des réseaux en matière de gestion des données de la recherche*. Ce travail rendait compte dans ses grandes lignes, des usages et des questionnements des réseaux sur la gestion des données, tout en apportant une vision des métiers transversale sur le sujet et les problématiques attenantes.

Il nous est apparu opportun d'aller plus loin et de détailler plus précisément les apports de nos réseaux métiers, compte tenu :

- des nombreuses actions de formation ou de sensibilisation mises en place,
- des compétences et expertises développées prenant appui sur des pratiques standardisées qui font leurs preuves sur le terrain,
- de la diffusion de recommandations et de solutions techniques et organisationnelles au sein des communautés grâce à la veille technologique et juridique réalisée très régulièrement.

Dans ce document, nous avons donc voulu témoigner des travaux réalisés au sein de nos réseaux métiers qui rendent compte de la gestion des données de la recherche tout en guidant le lecteur vers des bonnes pratiques en l'invitant aussi à cliquer sur des liens qui constituent des ressources lui permettant d'approfondir le sujet.

Ce guide est donc un document un peu hybride proche du vademécum, composé d'un inventaire d'actions de formations (conférences, séminaires, présentations), d'un répertoire de liens, de recommandations professionnelles, complétées de définitions utiles, pour approfondir le sujet de la gestion des données dans les réseaux métiers.

Il s'adresse à toute personne désireuse de se former à la gestion des données de la recherche, et son objectif est d'aider le lecteur à analyser son besoin et trouver des solutions parmi l'éventail des communications qui sont présentées. Il constitue aussi une invitation à se rapprocher des réseaux métiers.

Sommaire

Ce guide abordera l'ensemble des phases et actions nécessaires pour une gestion des données en accord avec les prérogatives de la science ouverte :

- Les deux premières phases 1 « *Imaginer et préparer* » et 2 « *Concevoir et planifier* », sont les étapes préparatoires d'un projet, où l'on se préoccupe d'avoir toutes les informations nécessaires à la bonne gestion des données et du projet. C'est l'étape où l'on réfléchit au plan de gestion de données, où l'on prépare les espaces de stockage et où l'on met en place les outils de gestion de projet. Cette partie, très générique et a pour objectif de conduire le lecteur à s'interroger sur ses besoins, les moyens dont il dispose, à se poser les bonnes questions et à s'orienter pour trouver des solutions adaptées dans un environnement riche, en construction et à surveiller.

Les phases suivantes apportent des éléments plus spécifiques au lecteur pour répondre à des besoins plus techniques

- L'étape 3 « *Collecter* » rend compte de la pratique de collecte et du processus d'acquisition des données (équipements, capteurs ...). Elle informe tout particulièrement sur l'usage des normes et standards d'interopérabilité nécessaires à la constitution des jeux de données pour les rendre Faciles à trouver, Accessibles, Interopérables, et Réutilisables (FAIR). Elle apporte aussi un éclairage sur les environnements de stockage des données existants et la nécessité de sauvegarder des données.
- L'étape 4 « *Traiter* » témoigne du prétraitement des données brutes acquises et collectées précédemment. Elle guide le lecteur sur la nécessaire préparation des fichiers de données pour les rendre ouverts et interopérables. La connaissance et la maîtrise des formats et standards est importante. Cette étape est également celle de l'organisation des données qui implique dans certains cas de développer des procédures d'intégration des données dans les bases de données ou d'utiliser un cadre d'application d'agrégation de données. Il est important

1. Anciennement DIST.

aussi à ce stade de se préoccuper du dépôt des données dans des plateformes de gestion locales² qui facilitent leur accès pour les scientifiques, et de mettre en place un contrôle qualité.

- la phase 5 « *Analyser* » est la phase d'analyse dans laquelle on s'occupe de définir et mettre en place des chaînes logicielles avec des méthodes et des outils. Cette partie informe le lecteur sur les plateformes, outils et méthodes utilisés principalement dans la communauté du calcul pour analyser et visualiser les données. Elle présente également quelques projets d'analyse sémantiques de données textuelles ainsi que des services Text and Data Mining.
- la phase 6 « *Préserver et archiver* » rend compte de l'importance de préserver et archiver les données sur le long terme. On s'attache dans cette partie à bien définir et clarifier les termes, réfléchir aux données pertinentes à préserver et voir quelles solutions s'offrent à nous.
- la phase 7 « *Publier et diffuser* » est la phase finale permettant de diffuser les données correctement à travers des catalogues de données, des thesaurus de mots clés fournissant une interopérabilité sémantique, des identifiants pérennes et des entrepôts de données, des datapapers.

Deux annexes reprennent des thèmes transverses à ces phases successives :

- l'annexe « *Infrastructures* » décrit le paysage des infrastructures destinées à la recherche scientifique et accessibles à la communauté scientifique qui travaille en France. Le recensement est loin d'être exhaustif mais a pour objectif de fournir un aperçu des différents types d'infrastructures et d'en présenter certaines plus en détail au travers des exposés réalisés lors d'événements organisés par les réseaux.
- l'annexe « *Reproductibilité* » rassemble des exposés concernant la reproductibilité et la répétabilité, sujets transverses à plusieurs phases. Comme dans l'ensemble de ce guide, les présentations référencées ont été réalisées dans le cadre de journées ou ateliers organisés par les réseaux.

Ce guide a été rédigé par quelques réseaux métiers et de ce fait n'a pas la prétention d'être exhaustif. Nous invitons donc d'autres réseaux à nous faire remonter leurs pratiques métiers dans le cadre de la gestion des données pour alimenter ce guide. Le lecteur est également invité à nous communiquer tous commentaires, remarques ou suggestions qui seraient de nature à améliorer et compléter ce travail en envoyant un mail sur [la liste de diffusion](#).

2. Il s'agit de plateformes développées en fonction des besoins par des unités ou établissements comme par exemple le data center du GRICAD qui propose aux équipes de recherche un espace de stockage et un service d'accompagnement pour la gestion des données.

Imaginer et préparer

Imaginer est la première étape de notre cycle de vie. C'est une phase préparatoire qui correspond à la connaissance et à l'identification des problématiques générales, techniques et juridiques associées à la gestion des données dans un projet de recherche ou dans la pratique quotidienne de nos métiers. Etape où l'on doit se projeter, s'informer, comprendre pour anticiper et envisager sereinement le déroulement d'un projet. C'est une étape initiale importante pour appréhender globalement la gestion des données, l'écosystème dans lequel elle s'inscrit avec ses contraintes et opportunités, les outils et infrastructures disponibles ou nécessaires, les politiques d'accompagnement et la multiplicité des acteurs qui interagissent, les réglementations en vigueur ou encore les compétences et expertises à acquérir.

L'apport des réseaux est ici important en termes de croisement des disciplines et des métiers pour apporter un éclairage global dans la nécessaire évolution des métiers et compétences et répondre au mieux aux besoins des communautés scientifiques.

1.1 Comprendre le paysage de la gestion des données

Avant d'aborder la gestion des données sous ses aspects techniques qui seront développés tout au long des étapes du cycle de vie de la donnée dans ce guide, nous souhaitons apporter une vision d'ensemble du paysage de la gestion des données. Ce paysage s'appréhende dans le cadre du mouvement open science, de la politique d'open data en particulier et par la connaissance de l'ensemble du processus de recherche depuis la compréhension des possibilités de financement de la recherche (attendu des financeurs H2020, ERC, ANR . . .) jusqu'à la diffusion, la valorisation et l'évaluation des résultats.

1.1.1 Connaître les politiques d'accompagnement des données au niveau européen et international

Différentes initiatives institutionnelles sont développées au sein de nos établissements en France ou à l'étranger pour accompagner la politique des données de la recherche. Il est intéressant de se pencher sur ces travaux pour anticiper les besoins et prévoir les évolutions stratégiques possibles au sein de notre environnement.

À l'occasion des « FréDoc 2013 », Simon Hodson (Directeur exécutif de CODATA) dresse un panorama très complet des différentes politiques institutionnelles, des tendances gouvernementales et internationales. Nous percevons très distinctement les défis et obstacles à lever pour la mise en place d'une gestion des données de la recherche. On comprend l'importance d'analyser le comportement des communautés de recherche pour parvenir à construire ensemble une politique autour des données et aussi l'intérêt d'une approche convergente « top down et bottom up » pour la mise en place d'actions de terrain qui rejoignent les actions de la gouvernance. Les nombreux défis à la mise en place d'une politique des données de la recherche sont bien présents à commencer par le fait d'instaurer au sein de nos communautés une culture du partage de la donnée et de mettre à disposition des chercheurs des infrastructures et des services de formation. Simon Hodson souligne aussi le rôle essentiel des politiques et parties prenantes pour mettre en place des actions et des concertations.

Les politiques d'accompagnement des données : une comparaison internationale

Simon Hodson, ISCU-CODATA, ANF « Frédocs2013 - Gestion et valorisation des données de la recherche », 2013, Aussois

En 2017 à l'occasion d'une ANF dédiée à l'organisation du management des données de la recherche et dans le contexte d'omniprésence du numérique et des défis sociétaux actuels, Francis André (Chargé des données de la recherche à la DIST du CNRS) présente l'évolution des pratiques scientifiques et le cadre stratégique offert par l'open science autour des données de la recherche.

On découvre à travers son intervention les principes FAIR et l'importance de disposer de métadonnées de qualité. On comprend également la nécessité du partage pour faire évoluer les connaissances. Francis André distingue dans sa présentation différents types de données, d'infrastructures et de services à l'échelle européenne et internationale avec un focus sur le fonctionnement et les groupes de travail de la Research Data Alliance (RDA). Il revient sur les résultats d'une enquête réalisée auprès des directeurs d'unité pour aborder la question du point de vue du chercheur et insiste sur la nécessité de réinventer nos métiers et de s'approprier la gestion des données.

Gestion des données de la recherche dans le contexte d'Open Science

Francis André, DIST-CNRSANF « Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données », 2017, Paris

La gestion des données s'organise également au sein des structures d'accompagnement au niveau européen comme en témoigne Susan Reilly (Directeur exécutif de la Ligue Européenne des Bibliothèques de Recherche LIBER) qui pointe en particulier les besoins de réorganisation et d'adaptation des structures d'accompagnement de la recherche pour aider les chercheurs dans la gestion des données. Le rôle des bibliothèques est crucial dans ces actions et les opportunités à saisir pour évoluer dans ce sens sont nombreuses. Partant du constat que sans une infrastructure qui aide les chercheurs à gérer leurs données de façon adéquate et efficace, aucune culture du partage n'est possible. Elle expose dans le cadre de LIBER une démarche proactive au sein des bibliothèques de recherche en Europe et présente 10 recommandations à suivre pour répondre aux besoins des chercheurs en termes de services et de supports.

From data management policy to implementation : opportunities and challenges for libraries

Susan Reilly, LiberANF « Frédocs2013 - Gestion et valorisation des données de la recherche », 2013, Aussois

On constate en effet aujourd'hui que les services communs de documentation, nouvellement investis de ces problématiques d'open acces et de gestion des données de la recherche ont entrepris une restructuration de fond au sein des Universités pour proposer des services d'accompagnement à destination des équipes de recherches (*Voir Arabesque n°95, ADBU, 2019*). De nouveaux services d'appui à la recherche se constituent peu à peu et proposent un soutien pour la publication en libre accès ou la rédaction de plan de gestion de données. (*voir SOS-PGD, répertoire des services opérationnels de soutien à la rédaction de plans de Gestion de données au sein des établissements de l'enseignement supérieur et de la recherche*)

L'INSU est aussi depuis les années 1990 à l'origine d'un dispositif d'accompagnement de la recherche à caractère national ou international qui bénéficie d'un processus de labellisation et qui a pour vocation d'apporter un service à la communauté scientifique. Il s'agit des **Services Nationaux d'Observation (SNO)** labellisés par la direction de l'INSU. Ces services ont été créés pour répondre au besoin de documenter sur le long terme la formation, l'évolution, la variabilité des systèmes astronomiques et des milieux terrestres, et de faire progresser les connaissances dans ces domaines.

1.1.2 Comprendre le contexte

Ces dernières années, la réglementation en matière de science ouverte a largement modifié le paysage des données de la recherche et a permis de mieux cadrer les pratiques scientifiques sur le plan juridique.

Un contexte politique favorable à la gestion et au partage des données

En 2018, à la suite des objectifs fixés par l'Europe, la France s'est dotée d'un **plan national pour la science ouverte**. Présenté le 4 juillet par Frédérique Vidal, la ministre de l'Enseignement et de la Recherche, ce plan prône la diffusion sans entraves des publications et des données de la recherche. Il comporte différentes mesures déclinées en trois axes stratégiques et pose les conditions du développement de la science ouverte dans nos établissements. Dans la continuité de ce plan, le CNRS, a rédigé une **feuille de route pour la science ouverte** s'appuyant sur des actions concrètes structurées autour de quatre grands objectifs : (i) 100% de la production scientifique en accès ouvert, (ii) développement d'une culture de la gestion et du partage des données, (iii) développement d'infrastructure pour la fouille et (iv) l'analyse des contenus et la transformation des modalités d'évaluation des chercheurs).

Enfin, le CNRS a publié en novembre 2020 un **plan Données de la recherche** avec l'objectif d'accélérer le développement vers la science ouverte, et d'encourager les chercheurs à rendre leur données accessibles et réutilisables. A côté de la mise en place d'une politique des données en phase avec les besoins des communautés scientifiques, ce plan envisage un nouveau mode de gouvernance et un plan d'action pour les données de la recherche.

Du côté des financeurs de la recherche, l'ANR dans son plan d'action 2020 réaffirme son engagement en faveur de la science ouverte. En lien avec le plan national pour la science ouverte, elle demande l'élaboration d'un Plan de Gestion des Données (PGD) pour les projets financés à partir de 2019. Partant des recommandations du **Comité pour la Science Ouverte (CoSO)**, elle a adopté un **modèle de PGD** proposé par Science Europe qui vise à harmoniser la gestion des données au niveau international. Ce plan constitue désormais un livrable de tout projet financé par l'ANR.

Le partage des données suppose également la mise en place d'un cadre juridique. La **loi pour la république numérique**, dite loi Lemaire, a posé ce cadre en octobre 2016 afin de favoriser l'ouverture et la circulation des données, de garantir un environnement numérique ouvert et respectueux de la vie privée et faciliter l'accès et la réutilisation des données. Le **Règlement général sur la protection des données (RGPD)** instaure quant à lui un nouveau cadre juridique pour la protection des données personnelles.

Un contexte technologique en constante évolution

Outre une attention particulière sur les besoins des communautés scientifiques, une veille technologique portant sur les services, outils, plateformes et infrastructures de stockage, de partage ou d'archivage des données de la recherche doit être assurée pour permettre d'adapter les moyens mis œuvre aux besoins exprimés.

De nombreuses initiatives internationales et programmes européens ont été constitués pour travailler à l'ouverture progressive des données de la recherche, produire et harmoniser des outils et méthodologies. Nombre d'entre eux sont des espaces de travail et d'échange pour les ingénieurs et chercheurs, comme **RDA (Research Data Alliance)** qui a pour motto « *Building the social and technical bridges to enable open sharing and re-use of data* » ou **Go FAIR** dont

l'objectif est de promouvoir les principes FAIR. Le « [European Open Science Cloud](#) » a été initié par la commission européenne. Il est défini comme « *The system resulting from the activities and initiatives promoted by the European Commission to support its policies on Open Science and Open Innovation 2.0* » (voir le [Glossaire de l'EOSC](#)).

Une présentation de Françoise Genova et de Francis André détaille le fonctionnement de la RDA, organisation internationale créée en mars 2013, pilotée par la communauté, qui vise à construire les ponts sociaux et techniques pour le meilleur partage des données. Les groupes d'intérêt et de travail y sont présentés par thématiques ainsi que les productions et recommandations issues de ces groupes. Un focus particulier est porté sur le nœud national RDA France, ses objectifs et ses activités.

Les activités de RDA : perspectives dans le cadre du nœud national français

Francis André, DIST-CNRS & Françoise Genova, Observatoire Astronomique de StrasbourgSIST 2018 : Séries Interopérables et Systèmes de Traitement, 2016, Guyancourt

Dans cette autre présentation, Volker Beckman (chargé de mission CNRS-EOSC et Directeur adjoint scientifique Calcul et Données IN2P3/CNRS) explique comment concrètement les chercheurs pourront utiliser l'EOSC. Il présente la stratégie européenne d'élaboration de ce Cloud européen lancé en 2018, qui coordonne les initiatives et projets de construction de cet espace à destination de la recherche et des chercheurs. Partant des nombreux projets qui ont été financés pour élaborer ce cloud ([EOSCpilot](#), [EOSC-Pillar](#), [EOSC-hub](#) etc.), il montre les possibilités de collaboration. Depuis 2019, une structuration est en cours avec la mise en place d'une gouvernance temporaire et une implication forte des ministères en charge de la recherche dans les différents pays européens. Ces travaux préparatoires devraient déboucher sur d'importantes opportunités de financement complémentaires dans le programme cadre « Horizon Europe ».

European Open Science Cloud (EOSC), opportunités pour la recherche en France

Volker Beckman, CNRS/IN2P3Rendre FAIR les données, mais quelles données préserver ? Atelier Dialogu'IST, 2020

Pour accompagner les communautés de chercheurs, une [feuille de route nationale des infrastructures de recherche](#) est mise à disposition sur son site par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Elle recense aujourd'hui 99 infrastructures et est régulièrement remise à jour.

Les projets de recherche au sein de ces infrastructures ont donné lieu à certains retours d'expériences qui témoignent de spécificités disciplinaires dans la gestion des données de la recherche. (Ils sont détaillés dans la partie dédiée aux [Infrastructures](#))

Il est important également de suivre attentivement l'évolution des espaces de partage des données de la recherche qui sont différents en fonction des communautés scientifiques. Les organismes de financement, les éditeurs ou les établissements de recherche ont pour coutume de recommander le dépôt des données dans des entrepôts, car ceux-ci permettent de conserver, rendre visible et accessible les données de recherche. Il en existe plusieurs catégories : entrepôts généralistes comme [Zenodo](#) ou [Dryad](#), institutionnels comme [Dataverse Cirad](#), [Datapartage](#) à INRAE, [dataSuds](#) à l'IRD ou thématiques comme [GBIF](#) pour les données de biodiversité, ou [Pangaea](#) pour les données des géosciences.

Des répertoires de données comme [Re3Data](#) (répertoire d'entrepôts créé par DataCite) ou [Cat OPIDoR](#) (catalogue de services dédié aux données de la recherche hébergé à l'INIST) sont accessibles pour guider les recherches.

Pour plus de détails, on se reportera à la section [Infrastructures](#).

1.2 Comprendre et respecter la législation en vigueur

Gérer les données de la recherche suppose de clarifier en amont les modalités de partage et de mise à disposition des données de la recherche et le cadre juridique applicable aux projets de recherche.

Comme le précisent les interventions de Nathalie Gandon (Frédocs 2018) ou Nathalie Le Ba (ANF Sciences des données), il existe un certain nombre de principes fondamentaux associés à la notion d'open data et un certain nombre de textes législatifs en France et en Europe qui réglementent ou impactent la gestion des données de la recherche et la réutilisation des informations publiques. Parmi ces textes, figurent principalement la [loi Valter](#) (2015) et la [Loi pour la république numérique](#) (2016) qui toutes deux élargissent le champ d'application de la [Loi CADA](#) et ont pour objectif de favoriser la réutilisation de l'information publique. La loi Valter instaure le principe de gratuité dans la réutilisation des informations publiques tandis que la Loi pour la république numérique (Loi Lemaire) conduit à l'obligation de mise en ligne spontanée des documents administratifs librement réutilisables (y compris à des fins commerciales). Ces deux lois sont à l'origine du principe d'ouverture ou d'open data par défaut. Les notions de « document administratif », d'universalité et de gratuité des informations publiques sont ici des notions incontournables à saisir pour passer d'une logique de demande citoyenne à une logique de diffusion volontaire des informations du secteur public.

Dans tous les cas, l'application des textes législatifs aux données de la recherche n'est pas toujours aisée. Nathalie Gandon, nous apporte des renseignements précieux à travers une check-list pour déterminer si les résultats de recherche sont ou ne sont pas des « documents administratifs » à diffuser. Il convient de s'interroger tout d'abord sur la nature et la forme du résultat concerné (le document doit être achevé), ensuite sur l'auteur du résultat (le document doit être produit dans le cadre d'une mission de service public) et enfin sur les conditions de production du résultat (collaboration publique ou privée). Il existe également de nombreuses exceptions prévues par la loi qui conduisent à une interdiction totale d'accès et de réutilisation (documents secret défense etc.). On trouvera sur ces supports le détail des exceptions liées aux données environnementales et personnelles.

Résultats de la recherche et open data : le cadre juridique

Nathalie Gandon, INRAANF « Fredocs 2018 - Démarches innovantes en IST : expérimenter, proposer, (se) réinventer », 2018, Albi

Questions juridiques autour de l'ouverture des données

Nathalie Le Ba (DAJ du CNRS)ANF « Sciences des données : un nouveau challenge pour les métiers liés aux bases de données », 2018, Sète

En complément de ces présentations, il est important de retenir que le droit des producteurs de bases de données (droit sui generis) est désormais « neutralisé » par la Loi sur la république numérique. Comme toute administration, les universités et établissements de recherche ne peuvent opposer leur droit de producteur de bases de données à la libre réutilisation des informations qu'elles produisent. Le principe d'ouverture par défaut s'applique. Pour plus d'information sur ce point nous vous invitons à consulter l'article de Lionel Maurel : [les universités françaises et l'Open Data après la loi numérique](#).

Les données à caractère personnel

La gestion des données implique également de porter un regard attentif à la législation sur les données à caractère personnel.

Les données personnelles, régies en France par la loi informatique et liberté (loi de 1978, modifiée le 20 juin 2018 pour adaptation au RGPD) font l'objet d'un traitement particulier. Entré en vigueur le 25 mai 2018 dans toute l'Union européenne, le [Règlement général sur la protection des données](#) (RGPD) instaure un nouveau cadre juridique pour la protection des données personnelles. Ce nouveau règlement renforce les droits des citoyens européens et responsabilise les organismes qui traitent les données pour garantir la protection des droits fondamentaux. Les principes énoncés dans ce règlement doivent être connus et respectés, car ils s'appliquent aussi aux activités de recherche. Le texte

prévoit néanmoins un régime spécifique, dérogatoire offrant une large marge de manœuvre aux chercheurs pour l'utilisation des données personnelles dans le cadre d'un projet de recherche. (Voir l'article de Lionel Maurel : [Données personnelles et recherche scientifique : quelle articulation dans le RGPD ?](#))

On trouvera dans l'intervention de Patrick Guillot (CIL des établissements universitaires de la ComUE Université Grenoble Alpes), une présentation riche et complète comprenant entre autres, un rappel des définitions et principes fondamentaux de la loi, une définition des « données à caractère personnel », un historique des principaux textes et un quiz de questions-réponses (vrai/faux) très utile pour comprendre l'évolution de la réglementation.

Prise en compte des données personnelles - Évolution de la réglementation

Patrick Guillot, Univ. Grenoble Alpes ANF « Traçabilité des activités de recherche et gestion des connaissances », Réseau Qualité en Recherche, 2017, Grenoble

L'INSHS a par ailleurs produit un guide pour la recherche « [Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte](#). Ce guide régulièrement mis à jour synthétise les règles applicables à chaque étape du cycle de vie des données et s'appuie sur des exemples concrets pour présenter des bonnes pratiques.

Les licences

Le choix des licences est également réglementé par la Loi pour une république numérique

L'ouverture et le partage des données impliquent par ailleurs l'utilisation de licences (GNU, [Creative Commons](#), [CeCILL](#), et autres) qui fixent les conditions dans lesquelles les données peuvent être réutilisées. La Loi pour une république numérique impose l'utilisation de licences applicables aux « informations publiques » d'une part et aux codes sources et logiciels d'autre part. Les licences utilisables sont référencées sur [cette liste](#) fixée par décret et applicable par les administrations. Si toutefois, le consortium du projet dans lequel les données ont été créées impose un autre choix, il faut s'y conformer. La [Licence ouverte / open licence](#) conçue par Etalab est considérée comme une licence de référence par le gouvernement pour la réutilisation et la publication de données publiques.

Pour accompagner les équipes dans le traitement juridique des données, un collectif de juristes issus de l'enseignement supérieur et de la recherche a rédigé un guide de référence « [Ouverture des données de recherche. Guide d'analyse du cadre juridique en France](#) » qui explique les principes à respecter en matière de diffusion des données.

1.3 Connaitre et comprendre les principes FAIR

Énoncés initialement par le groupe de travail FORCE 11, les principes FAIR « The FAIR Guiding Principles for scientific data management and stewardship » ont été publiés en mars 2016 dans la revue *Scientific Data*. Élaborés par des représentants du monde universitaire, de l'édition, de l'industrie et des organismes de financement, ils répondent aux besoins urgents d'amélioration des infrastructures permettant la réutilisation des données scientifiques.

1.3.1 Définir les principes FAIR pour guider les stratégies de gestion des données

Il s'agit d'un ensemble de [principes directeurs](#) visant à rendre les données de la recherche, Faciles à trouver, Accessibles, Interopérables et Réutilisables (FAIR) par les êtres humains et les machines. Ces principes 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. Il s'agit aussi en particulier de mettre l'accent sur le renforcement de la capacité des machines à rechercher et utiliser automatiquement les données afin de favoriser leur réutilisation par des particuliers.

Les principes FAIR ont pour objectif de guider le partage et la publication des données. Toutefois, s'il y a une volonté forte en faveur du partage et de la réutilisation des données (les principes sont adoptés par de plus en plus d'organismes

de financement de communautés scientifiques et sont également préconisés dans le plan national pour la science ouverte et dans la feuille de route du CNRS), il faut bien garder à l'esprit qu'appliquer les principes FAIR n'implique pas l'ouverture systématique des données. Le principe de base « aussi ouvert que possible, aussi fermé que nécessaire » reste en vigueur y compris lorsque l'on applique les principes FAIR.

1.3.2 Appliquer les principes FAIR - Retours d'expériences

Le groupe de travail inter-réseaux « Atelier données » s'est intéressé à l'application des principes FAIR. Une journée d'étude a été organisée en novembre 2018 avec l'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 a donné lieu à la production d'un [livret de synthèse](#).

Cette journée s'appuie plus spécifiquement sur les notions de pérennisation et d'interopérabilité des données dans les projets de recherche, et cherche à en comprendre les facteurs ressorts de réussite et les points sensibles à surveiller. Elle a pour ambition d'analyser les complémentarités des expériences au travers des métiers représentés par les réseaux ; de formuler des points de convergence de bonnes pratiques et d'accroître les échanges entre les réseaux de la MI sur des questions à forts enjeux pour l'évolution de nos métiers.

Interopérabilité et pérennisation des données de la recherche : Comment FAIR en pratique? Retour d'expérience

Atelier Données Inter réseau, 2018, Paris

1.4 Prévoir la traçabilité des données

Dans un environnement où l'information arrive en masse, pouvoir assurer la traçabilité des données est essentiel. Les données numériques représentent un enjeu majeur pour la recherche et il est important d'intégrer une démarche qualité au sein des structures de recherche pour disposer de données fiables et réutilisables.

Le réseau Qualité en Recherche particulièrement investi sur ce sujet, a élaboré en 2018, un guide de référence : [Traçabilité des activités de recherche et gestion des connaissances](#), à destination des agents des unités de recherche. Ce guide a pour 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 données de la recherche.

Alain Rivet, Responsable qualité et système d'information au CERMAV, illustre cette question à l'occasion de l'ANF Données 2016 en présentant la problématique de la donnée dans la perspective de la traçabilité des activités de recherche. Il pose la question du défi organisationnel de la gestion des données dans les laboratoires et les établissements face aux contraintes de plus en plus fortes des autorités de tutelles. Il souligne ainsi le besoin d'optimiser le fonctionnement de nos laboratoires, la solution étant de s'appuyer sur des référentiels comme la norme ISO 9001. La nécessaire confiance en la qualité d'une recherche suppose une maîtrise de l'ensemble des moyens d'acquisition, de traitement, de diffusion et de conservation des résultats.

Nos tutelles, en réponse à cette problématique d'intégrité scientifique, ont mis en place une stratégie nationale avec la rédaction début 2016 d'une charte de déontologie des métiers de la recherche qui insiste sur l'importance de permettre la traçabilité des travaux expérimentaux et la conservation des données de la recherche. Une bonne gestion des données de la recherche apparaît comme une réponse au problème soulevé.

Activités de recherche et gestion des connaissances

Alain Rivet, CNRS_CERMAVANF « Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données », 2016, Paris Voir aussi les vidéos : [La problématique de la donnée dans](#)

la perspective de la traçabilité des activités de recherche : contexte et enjeux (séquence 1) Le défi organisationnel de la gestion des données dans les laboratoires et les établissements (séquence 2) La qualité en recherche : contexte et définition (séquence 3) La qualité au service de la gestion des données dans les laboratoires (séquence 4) Conclusion : Développer les démarches « qualité » comme outil de gestion de données (séquence 5)

1.5 Envisager la curation des données

La curation des données est une activité essentielle dans la pratique de gestion des données, car elle assure la pérennité des données sur le long terme, leur qualité et leur réexploitation. Elle s'avère toutefois difficile à définir, car sa pratique se situe très souvent à la croisée de différentes disciplines. Elle s'applique tout au long du cycle de vie de la donnée et intègre des tâches de nature parfois différentes comme la sélection, la vérification, la normalisation ou encore l'enrichissement nécessaires à la publication des données.

« Les activités de curation de données permettent de faciliter la découverte et la récupération de données, de maintenir la qualité des données, de leur ajouter de la valeur et d'en fournir pour de futures réutilisations. Ce nouveau champ inclut la représentation, l'archivage, l'authentification, la gestion, la préservation, la récupération, et l'utilisation. »

Digital Humanities Data Curation

La définition ci-dessus semble de nature à mieux cadrer l'activité de curation pour la gestion des données de la recherche. Elle est proposée par le [Digital Curation Center \(DCC\)](#), une organisation britannique qui produit une expertise et fournit une aide pratique sur le stockage, la gestion, la protection et le partage des données de la recherche.

Le DCC propose également un Briefing paper « [What is Digital Curation](#) » qui explique les bénéfices d'une curation des données.

Pour illustrer une pratique de curation des données en SHS, Emmanuelle Morlock, Ingénieure au Laboratoire HiSoMa a présenté un travail réalisé dans l'univers de l'édition critique des sources.

Cette présentation s'organise en trois parties : les spécificités de la « data curation », les défis spécifiques aux SHS et les solutions proposées par l'encodage TEI (Text Encoding Initiative) de sources textuelle pour relever ces défis. Emmanuelle Morlock définit ici la notion de curation et les notions associées, les activités engendrées par cette activité et les défis qu'elles représentent pour les sciences humaines et sociales. Elle s'intéresse également aux types d'objets de la curation. Elle aborde ensuite le chapitre de l'édition savante qui l'amène à définir précisément ce qu'est l'édition numérique (un texte enrichi, exploitable par des machines) et à présenter, définir et expliquer le processus d'édition dans un format XML TEI. Elle explique aussi l'apport de la TEI dans la réponse aux défis posés par l'édition numérique (distinction de niveaux d'interprétation via le balisage, conservation et documentation des choix de manière formalisée) et termine sa présentation sur le rôle des « curateurs » pour repérer les manques dans un objectif de réutilisation à long terme ou pour aider les chercheurs à améliorer leurs pratiques de documentation de leurs données.

De quelques défis spécifiques de la curation numérique des données en SHS : petite incursion dans l'univers de l'édition critique de sources au format TEI

Emmanuelle Morlock, HiSoMa)Frédocs2013 - Gestion et valorisation des données de la recherche, 2013, Aussois

Dans le domaine spécifique des humanités numériques, il existe un guide auquel se référer : le [DH Curation Guide](#). Il est composé d'un recueil d'articles fiables sur la curation des données, contextualisés par des rédacteurs experts et des membres de la communauté. Ce guide, réalisé suite à une analyse de besoins exprimés par des professionnels en Humanités numériques dans le cadre d'un projet de recherche (Data Curation Education Program for the Humanities (DCEP-H) a été conçu avec l'objectif d'aider à relever les défis posés par la curation des données.

1.6 Prévoir l'archivage des données

La gestion des archives d'un laboratoire de recherche est une pratique assez peu courante au sein de nos unités, mais tend à se développer avec l'explosion du volume des données produites ou générées par les communautés de chercheurs.

Pour les archivistes de la section « Aurore » de l'association des archivistes français, « Les données de la recherche sont l'ensemble des informations et matériaux produits et reçus par des équipes de recherche et des chercheurs. Elles sont collectées et documentées à des fins de recherche scientifique. A ce titre, elles constituent une partie des archives de la recherche ».

Se préoccuper de l'archivage des données fait partie intégrante d'une bonne gestion des données. Dans une logique de préservation, l'archivage se conçoit très en amont d'un projet, dès la création de la donnée. Son objectif est de décrire, documenter, contextualiser les données pour pouvoir ensuite assurer leur diffusion et leur préservation à long terme. Il concerne tout type de données (bases de données, questionnaire d'enquête, données brutes, photos, etc.). Au-delà du stockage, il s'agit là de faire en sorte qu'une donnée soit réexploitable (intègre, lisible, intelligible) dans 10, 20 ou 50 ans par une nouvelle communauté de chercheur.

Les données sont des archives publiques dès lors qu'elles sont créées au sein d'un établissement public et l'archivage institutionnel est réglementé par la loi et notamment le [code du patrimoine](#). Les données doivent faire l'objet d'un tri, d'une sélection, idéalement à la suite d'un échange entre chercheur et archiviste en vue d'une conservation, si nécessaire, aux Archives nationales ou départementales.

Des outils existent pour aider à la sélection des données notamment le [référentiel de gestion des archives de la recherche](#). Ce référentiel est organisé par thématiques et indique pour chaque type de document sa durée de conservation, son sort final (tri, conservation, destruction) et les aspects légaux à connaître.

Pour plus de détails, on se reportera à la section *Préserver et archiver*.

1.7 Identifier les compétences et expertises pour la gestion des données de la recherche

Évoluer dans nos pratiques suppose de développer de nouvelles expertises et d'acquérir de nouvelles connaissances. Les réseaux professionnels, vecteurs de partage et d'échange sont particulièrement indiqués pour organiser et faciliter l'acquisition de nouvelles compétences.

1.7.1 S'informer et se former

La formation continue des personnels est fondamentale pour suivre l'évolution des métiers et des technologies.

Au CNRS, la formation continue est pilotée par le Service formation et Itinéraire Professionnel (SFIP). Celui-ci met en oeuvre des actions adaptées aux orientations et à la stratégie de l'établissement à travers deux dispositifs de formation principaux : les Actions Nationales de Formation (ANF) fortement orientée sur les technologies et ingénierie, et les « Écoles Thématiques » d'un contenu davantage scientifique et plutôt en relation avec les chercheurs. Le SFIP soutient également des actions régionales de formation.

Les réseaux métiers et réseaux technologiques

Dans ces dispositifs de formation institutionnels, *les réseaux métiers* sont fréquemment au coeur des propositions de programme, du montage et de l'organisation des ANF. Chaque année de nombreuses formations sont en effet régulièrement organisées par les réseaux, et les supports de formations présentés sont habituellement capitalisés sous une forme ou une autre (résumé, pdf, vidéo) sur les sites des réseaux.

Outre les ANF, les réseaux organisent également de manière autonome, sur budget propre attribués par la [Mission pour les Initiatives transverses et l'Interdisciplinarité](#) (MITI) ou par les Instituts du CNRS, des journées de séminaires qui

regroupent les membres des réseaux comme par exemple les [journées thématiques](#) organisées par le groupe de travail inter-réseaux « [Atelier données](#) » ou les [séminaires annuels](#) du réseau SIST de l'INSU.

Ils constituent bien évidemment des vecteurs importants de l'état de l'art et des connaissances à acquérir dans une discipline et contribuent à développer la connaissance d'un domaine de compétence.

Initiés et portés par des membres d'un même métier ou travaillant avec les mêmes technologies (outils, instruments, méthodes, etc.), les réseaux professionnels du CNRS ont vocation à faciliter les échanges d'informations et d'idées entre leurs membres.

Les réseaux favorisent le maintien et le développement des compétences, l'échange des pratiques professionnelles, l'implication et la motivation. Ils développent une connaissance fine de l'évolution des métiers et/ou des technologies de demain en assurant ainsi une veille pour les établissements d'Enseignement Supérieur et de la Recherche.

Les réseaux rattachés à la MITI du CNRS sont transversaux à tous les Instituts du CNRS, et accessibles aux personnels de l'Enseignement supérieur et de la Recherche,

La MITI accueille et pilote actuellement [22 réseaux](#) labellisés au sein de sa plateforme. Ils couvrent l'ensemble du territoire national et sont transverses à l'organisme.

Les réseaux labellisés par les instituts du CNRS viennent plus spécifiquement en support à leurs axes stratégiques scientifiques. Le blog RH du CNRS en recense un certain nombre dans son billet « [Evoluer, échanger, innover : les réseaux professionnels du CNRS](#) ».

Un dispositif de formation à distance sur les données de la recherche est accessible sur le site [DoRANum](#) (Données de la Recherche : Apprentissage NUMérique à la gestion et au partage). Cette plateforme met à disposition différentes ressources d'autoformation en libre accès sur la gestion et le partage des données de la recherche.

Le réseau national des [URFIST](#) (Unité Régionale de Formation à l'Information Scientifique et Technique), créé en 1982 est un réseau inter-académique structuré depuis 2017 en Groupement d'Intérêt Scientifique (GIS) qui a pour objectif de développer l'usage de l'IST dans l'enseignement supérieur et de la recherche.

Les sept unités régionales proposent chacune des ressources, documents pédagogiques ainsi que des formations (y compris doctorales) et manifestations scientifiques et professionnelles à [Bordeaux](#), [Lyon](#), [Paris](#), [Nice](#), [Rennes](#), [Strasbourg](#) et [Toulouse](#). Leur mission s'organise autour de trois axes principaux : la conception et la réalisation d'actions de formation, d'outils pédagogiques ainsi que la veille et la recherche dans le domaine des technologies de l'information.

Outre les actions de formation, d'expérimentations et innovations pédagogiques initiées par les Urfist, le réseau met à disposition un blog « [UrfistInfo](#) ».

1.7.2 Suivre les travaux du Collège « Compétences et formation du COSO »

Le Plan national pour la science ouverte, s'appuyant sur trois axes, engage les opérateurs de la recherche à « *mettre en œuvre une politique de science ouverte et à structurer et ouvrir les données de la recherche* ». La mise en œuvre de ce principe est exprimée à travers des objectifs et des actions de deux collèges du [Comité pour la science ouverte](#) (COSO) : les collèges [Données de la recherche](#) et [Compétences et formation](#).

Les collèges sont des groupes d'experts (plus de 200 à l'heure actuelle) qui impulsent et mettent en œuvre les projets en s'appuyant sur les acteurs, notamment ceux de la formation à la science ouverte.

Parmi les réalisations récentes, on peut citer le guide « [Pour une politique des données de la recherche : guide stratégique](#) » où le CoSO émet sept recommandations pour aider à la formalisation et à la mise en œuvre d'une politique des données de la recherche au sein des établissements de l'ESR, la réalisation du [Passeport pour la science ouverte](#) destiné aux doctorants de toutes disciplines, le guide « [Je publie, quels sont mes droits](#) » qui répond aux questions que se posent le plus souvent les auteurs de publications scientifiques sur leurs droits et la contribution à la session 3 du Mooc « [Recherche reproductible : principes méthodologiques pour une science transparente](#) ».

Concevoir et planifier

Dans cette étape du cycle de vie de la donnée, il s'agit de définir les tâches à accomplir pour réaliser le projet de recherche, d'élaborer un planning, de rechercher d'éventuels partenaires et financements, d'élaborer les spécifications nécessaires (i.e. de définir précisément les éléments fonctionnels et techniques souhaités), de définir les données et les métadonnées qui seront utiles, de penser au futur plan de diffusion et bien d'autres actions de préparation et de planification.

Pour ces travaux de conception et de planification, les réseaux apportent un appui sur la gestion de projet, les méthodologies de conduite de projet qui permettent par exemple la définition des indicateurs utiles au projet, les outils pour assurer l'interopérabilité des systèmes mis en oeuvre.

Ils fournissent des recommandations et des retours d'expérience pour la rédaction des plans de gestion de données, pour la définition du type de données à collecter, l'identification de nouveaux supports de publication. . .

À ce stade, il est aussi nécessaire de prévoir le mode de collecte et de stockage afin d'organiser la traçabilité en amont, traçabilité qui permettra de garantir la réutilisation des données.

2.1 Évaluer les besoins liés au projet

Lors du montage du projet ou au plus tard en début de projet, il est nécessaire d'évaluer les différents besoins, de mettre en place une organisation et les outils collaboratifs nécessaires à son bon déroulement. Les paragraphes suivants abordent différents points de vue complémentaires.

2.1.1 Anticiper les interfaçages nécessaires (avec les utilisateurs ou entre bases de données)

Il convient de prendre en compte les besoins des utilisateurs du projet, quelquefois appelés « use cases », et les besoins entre les différents systèmes qui seront sollicités, qu'ils existent déjà ou qu'ils soient élaborés pour le projet.

À l'occasion de l'ANF « Système d'information embarqué, cahier/carnet de terrain et de laboratoire électronique : quelles interactions avec les bases de données ? » organisée en mai 2016, Nadine Mandran, du laboratoire LIG, explique comment intégrer l'utilisateur au sein d'un projet. Elle présente en parallèle, la méthode Agile et la Démarche centrée utilisateur. Le contenu de cette présentation est très concret et illustré par des exemples.

Méthodes pour intégrer l'utilisateur dans la construction des applications

Vidéo : Nadine MANDRAN, LIG Séminaire « Système d'information embarqué, cahier/carnet de terrain et de laboratoire électronique : quelles interactions avec les bases de données ? », 2016, rBDD

2.2 Mettre en place une gestion de projet

Les aspects de conception et de planification nécessitent de mettre en place une méthodologie de gestion de projet. En complément, une analyse des risques et une analyse SWOT (méthode permettant d'analyser les forces, les faiblesses, les opportunités et menaces liées à un projet) pourront également être menées.

De façon à mieux comprendre l'amplitude de la gestion de projet et sa nécessaire adéquation avec la thématique scientifique du laboratoire et les contraintes qui lui sont afférentes, on pourra consulter la présentation de Myriam Ferro réalisée en 2017. Elle présente la démarche qualité mise en place au laboratoire « Étude de la dynamique des protéomes » ainsi que les outils déployés pour appuyer la démarche.

Gestion de projet dans le domaine de la recherche en biologie avec la mise en place d'outils tels que la mise en place d'une procédure R&D, de fiche projet, l'utilisation d'un LIMS et les gestions des anomalies

Myriam Ferro, U1038 BGE CEA/Inserm UGAR Rencontres du Réseau Qualité en Recherche : Traçabilité des activités de recherche et gestion des connaissances, Grenoble, 2017

Dans d'autres domaines, on pourra s'inspirer de la « Méthode de conduite de la recherche en informatique centrée humain ». Il s'agit des domaines de la recherche en informatique qui intègrent des utilisateurs pour construire de la connaissance scientifique et des outils supports à cette recherche. À titre d'exemple, nous pouvons citer les domaines concernés comme le domaine des Systèmes d'Information (SI), de l'Ingénierie des Interfaces Homme-Machine (IIHM) ou celui des Environnements Informatiques pour l'Apprentissage Humain (EIAH). Dans ces travaux de recherche se pose le problème du processus de conduite de la recherche et de la traçabilité des résultats et de la qualité des données. Cette méthode offre des outils conceptuels et techniques pour garantir la traçabilité du processus de conduite de la recherche. Elle porte le nom de THEDRE pour « Traceable Human Experiment Design Research ».

THEDRE : Méthode de conduite de la recherche en informatique centrée humain

Nadine Mandran, LIGRencontres du Réseau Qualité en Recherche : Traçabilité des activités de recherche et gestion des connaissances, Grenoble, 2017

À l'échelle d'un organisme, il est aussi possible de doter les équipes de méthodes et d'outils de gestion de la qualité. L'INRA a développé une nouvelle politique qualité et en appui à cette politique, un outil de diagnostic, EureQUA : une méthode pour manager tout type d'activité. Diane Briard, lors de l'ANF qualité 2019, présente cette politique et détaille le fonctionnement d'EureQUA.

Outil de Diagnostic EureQUA : une méthode simple d'aide à la décision en appui au pilotage des activités de recherche et d'expérimentation

Diane Briard, INRAANF Qualité : Faire simple et utile (QUALSIMP), Nancy, 2019

La gestion d'un projet inclut aussi l'assurance produit. Il s'agit de l'ensemble des dispositions et activités définies et mises en place pour garantir que le produit atteigne les objectifs définis dans le cadre d'un projet ou d'une mission et qu'il soit sûr, fiable et disponible. C'est avant tout une question de bon sens et d'organisation interne. L'assurance produit s'applique à tout type de projet de manière transverse sur toutes les thématiques techniques et interagit avec tous les acteurs du projet. Ces activités couvrent :

- la maîtrise des risques et la sûreté de fonctionnement ;
- l'assurance qualité (en conception et fabrication, approvisionnement et gestion de la sous-traitance, gestion des équipements, traçabilité) ;
- la maîtrise de la qualification des matériaux, composants et procédés ;
- la maîtrise et le contrôle de la contamination particulaire, moléculaire... ;
- la gestion de la documentation et de la configuration ;
- l'assurance qualité Logiciel.

Le réseau Qualité en Recherche a élaboré une [guide assurance produit](#) qui détaille ces points.

2.2.1 Analyser les risques

La sécurité de l'information est définie comme la « protection de la confidentialité, de l'intégrité et de la disponibilité de l'information ». Elle devient aujourd'hui une des problématiques majeures de nos unités.

Forts de ce constat, nous devons envisager la finalité de « protection du patrimoine scientifique » à travers des enjeux principaux :

- garantir la disponibilité de l'outil de travail pour l'ensemble des personnels de la structure
- garantir la confidentialité des informations, qu'elles soient professionnelles ou personnelles ;
- garantir l'intégrité des informations et des personnes ;
- assurer la protection des données à caractère personnel et / ou sensibles collectées, produites ou gérées par la structure (données scientifiques et techniques, données de gestion administrative, données individuelles) ;
- assurer la protection juridique (risques administratifs, risques pénaux, perte d'image de marque).

Une analyse de risque telle qu'évoquée dans le [Guide des bonnes pratiques pour les Administrateurs Systèmes et Réseaux](#) apparaît comme une réponse aux besoins de protection des données de nos unités de recherche.

L'analyse de risque permet alors d'identifier les objectifs de sécurité et les mesures à prendre adaptées aux besoins de protection de données de l'unité. Elle sert d'élément à l'élaboration de la politique de sécurité du Système d'Information (PSSI).

L'analyse de risques et la gestion des risques sont des processus importants de la gestion de projet. Un ensemble de présentations réalisées en décembre 2015 en dresse un panorama.

L'analyse des risques, leviers de différenciations des projets et d'amélioration de la qualité de la recherche

Thierry Bontems, UMR Pacte, Professeur affilié à Grenoble École de Management6e rencontre du réseau Qualité en Recherche, Biarritz, 2015

Le management des risques dans un projet

Marielle LACOMBE, Observatoire Midi Pyrénées, Toulouse6e rencontre du réseau Qualité en Recherche, Biarritz, 2015

La cartographie des risques pour améliorer les services relatifs à la gestion des contrats et conventions

Catherine ROCH – Sabine GOULIN, Université de Lorraine6e rencontre du réseau Qualité en Recherche, Biarritz, 2015

Présentation de l'analyse SWOT : les usages et les conditions d'emploi de la méthode

Sabine GOULIN, Université de Lorraine6e rencontre du réseau Qualité en Recherche, Biarritz, 2015

Définir le risque associé à un jeu de données

Eric Quinton aborde la question de la protection des données dans un contexte de menaces informatiques. Que représente la donnée ? Il précise qu'une donnée n'a pas de valeur intrinsèque, c'est la représentation d'une réalité, elle dépend de son contexte d'acquisition, de son traitement. Lorsque l'on travaille sur les données, on travaille toujours sur un processus d'acquisition dans le cadre d'un référentiel. Certaines données doivent être protégées, car elles comportent un risque si elles venaient à être diffusées et réutilisées. Face aux menaces informatiques qui n'ont jamais été aussi nombreuses (piratages, attaques, arnaques), il importe de définir le risque et de comprendre comment l'intégrer à la gestion des données. Il existe plusieurs définitions, mais l'on s'accorde pour dire que c'est la conjonction entre une cible, un impact, une cause et une probabilité. Une fois la cible définie (un jeu de donnée) il convient de définir l'impact ou la gravité. Celui-ci s'évalue selon trois critères (confidentialité de l'information, intégrité des données et disponibilité du système) classés chez IRSTEA selon un schéma détaillé correspondant à une échelle de 1 à 4. Il est important aussi de calculer l'impact en cas de défaillance – pour chaque critère, on note 4 niveaux d'impact estimés selon 4 thématiques. L'étude débouche sur un tableau récapitulatif qui est reporté dans le plan de gestion de données. Ce tableau permet de définir l'impact maximal et la sensibilité du système. On ne peut connaître toutes les menaces et les causes à prendre en compte. L'usage est de consulter les recueils de bonnes pratiques et de se référer aux listes et référentiels existants. La probabilité d'occurrence d'une menace est à évaluer en fonction du risque associé – on note trois niveaux de risque : opportuniste, ciblé, attaque concertée. En conclusion, Eric Quinton explique comment intégrer concrètement le risque dans le plan de gestion des données en présentant des exercices pratiques.

Définir le risque associé à un jeu de données

Eric Quinton, IRSTEA Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données, 6-8 juillet 2016, Paris Voir aussi les vidéos : La protection des données dans un contexte de menaces informatiques (séquence 1) Qu'est-ce que le risque et comment l'intégrer à la gestion des données (séquence 2) Conclusion : Comment intégrer concrètement le risque dans un plan de gestion des données : exercices pratiques (séquence 3)

2.2.2 Sauvegarder les données

À ce stade, il est important de prévoir une sauvegarde pour les données qui seront collectées, créées, construites dans le cadre du projet. La sauvegarde des données est un point important de la gestion du projet qui répond au risque de perte des données, c'est aussi un point qui doit être traité dans le plan de gestion de données traité au paragraphe suivant. Étant donnée l'importance de la sauvegarde des données, nous préférons, au risque de redondance, mettre l'accent ici sur cette nécessité. L'objet de cette phase préparatoire n'est pas de discuter le choix d'une technologie ou d'une stratégie, mais simplement de se préoccuper de prévoir la sauvegarde des données, et se poser les bonnes questions :

- Quel volume approximatif devons-nous sauvegarder ?
- Selon quelle périodicité : quotidienne ? hebdomadaire ? mensuelle ?
- Les baies de stockages sont-elles disponibles ?
- Sont-elles sous contrat de maintenance ? Y a-t-il besoin d'une externalisation ?
- Les données devront-elles être accédées fréquemment ? en temps réel ?
- Les infrastructures informatiques ont-elles suffisamment d'espace de stockage disponible ?
- Etc. . .

2.3 Amorcer un plan de gestion de données

Un Plan de Gestion de Données (PGD), ou Data Management Plan (DMP) en anglais, est un document formalisé - un livrable du projet pour la plupart des appels à projets actuellement - explicitant la manière dont seront obtenues, documentées, analysées, disséminées et utilisées les données produites au cours et à l'issue d'un processus ou d'un projet de recherche.

À noter qu'il existe des modèles de plans de gestion de données dits « de structure » dont la période considérée s'étend au-delà de la durée d'un seul projet. Ce type de modèle s'applique par exemple aux plateformes et est donc sans doute de façon générale plus adapté aux besoins du personnel technique.

PGD de structure contrainte ou opportunité ?

P. Clastre, URFM, INRA PACAMontpellier Science Ouverte : Plan de Gestion de Données, 2019

L'initialisation du plan de gestion de données dans cette phase est un préalable à sa mise à jour nécessaire dans les étapes suivantes. Le PGD doit suivre les évolutions du projet.

2.3.1 Comment créer un plan de gestion des données ?

Cette présentation de Marie-Claude Quidoz présente dans une première partie le plan de gestion de données en perspective du cycle de vie des données et détaille les principes FAIR. Une seconde partie présente différents modèles de plans issus de plusieurs origines (INRAe, appels à projets). C'est une excellente entrée en matière pour comprendre rapidement ce qu'est concrètement un plan de gestion de données.

Plan de gestion des données.

Marie-Claude QUIDOZ, CEFE/CNRSSemaine TEMPO, Sète, 2019

Cette présentation de Marie Puren a été conçue pour animer un atelier de formation qui avait pour objectif de définir un plan de gestion de données, identifier les éléments clés qui le constituent et le créer. Cette présentation contient tout d'abord des éléments propres à définir les données de la recherche, le modèle d'ouverture dans lequel elles s'inscrivent, les initiatives européennes et nationales qui les soutiennent. Elle focalise ensuite sur la pratique de la gestion à proprement dite des données et ses implications (gérer, stocker, déposer), mais surtout définit, décrit le contenu formel du Plan de Gestion de données et les différentes étapes de gestion. Elle présente concrètement sa

structuration (description des données, standards et métadonnées, partage et archivage des données), elle aborde les questions juridiques et les bonnes pratiques de gestion notamment le FAIR DATA.

Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données

Marie Puren, INRIAANF Participer à l'organisation du management des données de la recherche, gestion de contenu et documentation des données, 2017, Vandoeuvre-lès-Nancy

Comme l'ont introduit les présentations précédentes, un plan de gestion de données ne se rédige pas seul, mais au contraire en associant les différents acteurs du projet. Il s'agit donc de rédiger collaborativement le PGD. La plateforme « DMP OPIDoR » de l'INIST fournit un service en réponse à ce besoin à l'ensemble de la communauté enseignement supérieur recherche en France pour rédiger de façon collaborative un PGD. Après un rappel du contexte autour de la gestion des données, cette présentation montre avec de nombreuses copies d'écran comment utiliser « DMP OPIDoR ». À ne pas manquer si vous souhaitez un panorama du contexte et si vous souhaitez savoir comment réaliser en pratique un plan de gestion de données en collaborant avec vos collègues.

Data management plan ? Plan de gestion de données ? DMP OPIDoR vous guide !

Laurent RASSINOX, Marie-Christine JACQUEMOT-PERBAL, INIST,SIST 2018 : Séries Interopérables et Systèmes de Traitement, Guyancourt, 2018

Des discussions, lors des Journées Calcul et Données 2019 (JCAD 2019), autour d'une Table ronde intitulée « Les Plans de Gestion des Données des projets Scientifiques, quels impacts pour les centres de Calcul et de Données ? » vous permettront aussi d'en voir les implications pour les centres de calcul et de données.

Vidéo de la table ronde

Animation : Nicolas Renon, CALMIP et participants : Windpouire-Esther Dzale-Yeumo, Emmanuel Courcelle, Jean-Yves Nief, Jean-Philippe Proux, Geneviève RomierJCAD 2019

2.3.2 Créer un plan de gestion de logiciel

Les logiciels sont aussi des données, un peu particulières et qui méritent donc un modèle approprié de plan de gestion : le plan de gestion de logiciel. Le projet PRESOFT propose un modèle adapté à la fois au logiciel et au contexte de la recherche en France. Après une présentation de ce contexte, du modèle et de la procédure associée, les apports de PRESOFT sont détaillés. À noter que le modèle proposé par PRESOFT s'étend sur l'ensemble de la « vie » du logiciel depuis l'idée, avec les documents préparatoires, jusqu'à la préservation (sous toutes ses formes) et qu'il prend en compte toutes les formes de financement (projets, stages. . .). Le modèle est disponible sur DMP OPIDoR.

Plans de gestion de logiciels

Geneviève Romier, Vincent Breton, CNRS-IN2P3JCAD 2019

2.3.3 Retour d'expérience

Afin de conclure ce tour d'horizon des plans de gestion de données, ce retour d'expérience relatif au domaine de la biodiversité vous permettra de mieux comprendre comment utiliser les plans de gestion de données comme un véritable outil de gestion qui va bien au-delà du document administratif nécessaire à la validation du projet.

Du Plan de Gestion des Données au Datapaper : suivi des données scientifiques tout au long de leur cycle de vie.

HEINTZ, Wilfried, INRA DynaforSIST 2018

Enfin, à des fins pédagogiques, le CEFE a rédigé un [plan de gestion de données](#) sur un projet fictif de « Suivi de population de poissons dans le lac du Bourget ». Il correspond à la version qui devra être transmise dans les 6 mois qui suivent le démarrage scientifique du projet financé par l'ANR. Il est disponible dans la rubrique [DMPs publics](#) de l'outil DMP Opidor.

2.4 Identifier les infrastructures adaptées au projet : fournisseur du service, fonctionnalités, capacités et services

Une fois les besoins exprimés, il faut identifier les infrastructures nécessaires à la réalisation du projet. On apportera également un soin particulier aux différents critères de choix de ces infrastructures.

La section *Infrastructures* de ce guide présente différents types d'infrastructures destinées à la recherche que ce soit au niveau européen, national, régional ou par thématique scientifique.

Les modes de collecte et de stockage sont détaillés dans la section collecter de même que l'utilisation d'un cahier de laboratoire.

2.4.1 Les bases de données

Les bases de données font partie des différents types de services pouvant être utilisés dans le cadre d'un projet et qui sont proposés par les infrastructures ou disponibles au sein des laboratoires. Plusieurs types de bases de données sont disponibles (SQL, noSQL, ...).

La gestion des données dans des bases de données relationnelles est un gage de structuration cohérente, et permet une interrogation des données par des opérateurs du langage SQL (System Query Language)

Plusieurs Systèmes de Gestion de Bases de Données (SGBD) existent dans le monde du logiciel libre, cependant, PostgreSQL est le SGBD conseillé par la [circulaire Ayrault](#) que Marie-Claude Quidoz a présentée en 2017.

La maîtrise de PostgreSQL est donc importante et plusieurs formations complètes ont été organisées à ce sujet :

- ANF « PostgreSQL Administration, premier niveau », 2019
- ANF « PostgreSQL Performance », 2019

Bien aborder la mise en place est l'objectif de la présentation « Comment concevoir une base de données en archéométrie ? », réalisée en juin 2014 par Isabelle BALY et Philippe GRISON. Ils présentent les différentes étapes nécessaires à la conception et à la réalisation d'une base de données en archéométrie. Ils en détaillent les différentes phases : analyse ou d'audit, modélisation & développement de la base, migration des données et déploiement & développement d'un SGBD.

Chaîne opératoire de réalisation d'une base de données.

Isabelle BALY et Philippe GRISON, Comment concevoir une base de données en archéométrie, 2014.

Cette autre présentation de Marie-Claude Quidoz s'intéresse à la problématique de la traçabilité des données appliquée cette fois aux bases de données. Elle recommande en particulier, dès que l'on traite des données de noter toutes les opérations faites (insertion / suppression / modification). Pour cela un mécanisme d'historisation doit être mis en place au moment de la création de la structure de la base de données. Cette historisation repose sur un mécanisme de déclencheur qui s'active sur les actions cités précédemment. Dans le cadre de PostgreSQL, le logiciel de SGBD recommandé dans la [circulaire Ayrault](#), cette historisation peut être automatisée grâce à l'extension E-Maj. La mise en place de ce mécanisme permet aussi d'envisager de pouvoir rejouer une requête et de reproduire le résultat tel qu'il était quand, par exemple, un identifiant (un DOI ou autre identifiant) a été défini.

Présentation générale sur la problématique de la traçabilité des données appliquée aux bases de données

Marie-Claude Quidoz, CEFEAtelier Traçabilité, 2018

Cette dernière présentation synthétique décrit les différentes facettes de la traçabilité d'un jeu de données, elle permet de découvrir l'extension E-Maj citée au paragraphe précédent. Cet outil, utilisable avec PostgreSQL, sous licence GPL, composé d'un client web et de l'extension PostgreSQL, permet de déplacer des contenus de données dans le temps avec une granularité de niveau de table. E-Maj permet également de dénombrer, consulter, annuler et rejouer des ensembles de tables applicatives en enregistrant les mises à jour.

E-Maj comme « Enregistrement des Mises A Jour » : Et vos données PostgreSQL voyagent dans le temps ! Un cas d'utilisation pour tracer les données PostgreSQL et E-Maj par la pratique

Marie-Claude QUIDOZ, Philippe BEAUDOIN, 2018ANF « Sciences des données : un nouveau challenge pour les métiers liés aux bases de données », 2018, Sète

Enfin, il peut être souhaitable de créer un Identifiant Universel Unique (Universally unique identifier - UUID) avec PostgreSQL. Nicolas Raidelet explique comment faire dans cette présentation.

UUID avec PostgreSQL : Pourquoi ? Comment ?

Nicolas Raidelet, IrsteaWebcast RDBB, 2017

2.4.2 La gestion des collections

Collec-Science est un logiciel web qui a été créé pour suivre les échantillons collectés lors des campagnes d'acquisition, et permet de répondre, entre autres, à ces questions :

- Où est stocké l'échantillon ?
- D'où vient-il, quelle est sa généalogie (protocole de collecte, métadonnées associées à l'échantillon et ceux de ces ancêtres) ?
- Quelles transformations ou opérations a-t-il subies ?
- Sous quelle forme est-il conservé, existe-t-il un risque à le manipuler ?

Fruit d'une collaboration initiale entre Irstea (centre de Bordeaux), le laboratoire Epc à Bordeaux, le LIENS à La Rochelle, il a été enrichi avec la participation de nombreux autres laboratoires, dont les laboratoires Chrono-environnement à Besançon, Edytem à l'Université Savoie - Mont Blanc, etc. Il a été choisi par le Réseau des Zones Ateliers pour assurer le suivi des échantillons.

Stockez et retrouvez vos échantillons avec Collec-Science

Marie-Claude Quidoz site web RBDD, 2018

Cette phase du cycle de vie de la donnée concerne les aspects d'acquisition et de collecte des données ainsi que la constitution des jeux de données (« dataset » en anglais) avec leurs métadonnées descriptives. Il s'agit donc, dans cette phase, de travailler sur les processus d'acquisition des données qui peuvent être obtenues au moyen de divers médias selon le domaine étudié : capteurs environnementaux, instruments, sondages, modèles numériques... Une fois les données acquises, il est nécessaire et indispensable dans l'objectif de les rendre « FAIR », de les décrire avec leurs métadonnées associées.

La description de ces jeux de données nécessite d'utiliser, autant que faire se peut, des référentiels de vocabulaires contrôlés (thésaurus) si possible standardisés et les plus appropriés au domaine étudié. Il est conseillé de gérer les jeux de données dans un environnement technique qui permette d'assurer la sauvegarde, l'archivage, le « versionning », l'accessibilité et l'interopérabilité des données. Cette gestion se fait via des infrastructures techniques, des bases ou des supports qui doivent être fiables et bien documentés, et ce dans le respect des règles de traitement spécifiques des données personnelles.

Cette phase « Collecter » va nécessiter :

- de disposer des données et de fournir les métadonnées nécessaires pour apporter toutes les informations utiles à la description des données brutes elles-mêmes (libellés des paramètres, unités de mesure, localisation, propriétaires etc.), ainsi que sur les dispositifs d'acquisition (capteurs de mesures, modèles numériques...);
- de mettre en place des chaînes de collecte : du capteur jusqu'aux espaces disques et aux applications sur des serveurs où les traitements pourront être réalisés, avec la documentation adaptée;
- d'utiliser des protocoles si possibles normalisés ou standardisés pour présenter les données brutes et les dispositifs d'acquisition (capteurs...) et les rendre interopérables;
- de mettre en place une gestion et conduite de projets pour faire travailler ensemble les différents acteurs intervenant dans la chaîne de collecte : électroniciens, informaticiens, chercheurs...;
- de disposer de cahier de laboratoire, tablettes de terrain ou supports divers pour consigner les relevés et métadonnées observées;
- de définir le stockage nécessaire à la collecte de données : travailler en amont avec une équipe informatique en mode projet (gestion de projet).

3.1 Utiliser des normes et des standards d'interopérabilité

L'Association Francophone des Utilisateurs de Logiciels Libres (AFUL) donne une définition de l'**interopérabilité** qui est « la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs, et ce sans restriction d'accès ou de mise en œuvre ». Développer l'interopérabilité consiste donc à mettre en place et utiliser des normes et des standards qui fixent des règles permettant d'assurer le bon fonctionnement et les échanges entre deux systèmes informatiques.

Appliquée aux données, l'interopérabilité permet de rendre les données accessibles et réutilisables. Pour parvenir à cela, il est nécessaire d'utiliser des protocoles d'accès et des formats des données « ouverts », normés ou standardisés, d'une part, au niveau des formats de fichiers et d'autre part, au sein des outils informatiques qui serviront à échanger, diffuser et lire les données.

3.1.1 Les standards de métadonnées

Dans l'optique d'une gestion « FAIR » des données, il est nécessaire, dans la mesure du possible, de suivre des normes et des standards pour la description des métadonnées, les formats de fichiers et les protocoles d'échange de données.

Catherine Morel-Pair propose une présentation riche et complète sur les formats et métadonnées qu'elle détaille de manière très approfondie et restitue dans le cadre de leur utilisation pour la gestion de contenu et la documentation des données. Elle aborde en introduction les notions de données de la recherche, de données FAIR, d'interopérabilité et de Data Management Plan.

- la première partie de sa présentation porte sur les fichiers de données (organisation et nommage, format et critères d'interopérabilité-pérennité)
- la deuxième partie est dédiée aux métadonnées et à la documentation (définitions, présentation des standards, des identifiants pérennes pour les données et syntaxes d'échange). Elle termine par un focus sur les sites de dépôt, de portails ou d'entrepôts de données et leur schéma de métadonnées associées.

Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données

Vidéo : Catherine Morel-Pair , INIST, CNRSANF « Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données » - 2016 Paris

Les métadonnées dans un Plan de Gestion de données

Cette présentation de Marie Puren a été conçue pour animer un atelier de formation qui avait pour objectif de définir et comprendre l'importance des métadonnées dans le cadre de la rédaction d'un DMP. Elle définit, en donnant des exemples, ce qu'est une métadonnée, à quoi elle sert, quelle information elle donne. Elle distingue et détaille la spécificité des métadonnées de description, des métadonnées de gestion et des métadonnées de préservation. Elle aborde ensuite le chapitre du cycle de vie des métadonnées (créer, entretenir, mettre à jour, stocker, gérer la suppression des données, publier). Elle spécifie les métadonnées à faire figurer dans un DMP, explique comment les collecter et propose quelques outils d'extraction automatique de métadonnées. Autour de la notion de métadonnée, elle précise l'importance de définir des responsabilités en s'appuyant sur les chercheurs, documentalistes, bibliothécaires et informaticiens. Elle complète sa présentation avec une description des principaux standards interdisciplinaires et disciplinaires de métadonnées. Elle explique où et comment choisir ces standards. Elle explique également l'intérêt d'associer des ontologies ou vocabulaires contrôlés. Les dernières recommandations de sa présentation portent sur la gestion des métadonnées à long terme, l'importance d'évaluer leur qualité et revient sur la notion d'ouverture des métadonnées et la nécessité de choisir des licences pour nos métadonnées.

Les métadonnées dans un DMP

Marie Puren, INRIAANF « Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données », réseau Renatis, Paris, 2017

3.1.2 Les référentiels de métadonnées

Les référentiels de métadonnées peuvent être des standards ou des normes ; ce sont des documents importants qui se chargent de définir les informations nécessaires pour décrire les données elles-mêmes. De ce fait, ils sont utilisés pour donner toutes les informations nécessaires à la compréhension et à l'utilisation des données et ainsi faciliter leur réutilisation. Il est donc fortement recommandé de décrire ses données avec des normes ou des standards reconnus dans les disciplines concernées. Le choix d'un standard de métadonnées va dépendre du type de ressource, du domaine d'application, mais également de la communauté à laquelle on s'adresse.

À cet effet, le site du [Digital Curation Centre](#) recense les standards de métadonnées par grandes disciplines (biologie, physique, sciences sociales, sciences de la terre...). Des outils informatiques, permettant de passer d'un standard à un autre, sont également disponibles.

On trouve plusieurs standards et normes qui permettent de définir un ensemble de métadonnées sur des jeux de données. Parmi les standards les plus connus ou utilisés, citons le Dublin Core qui est un standard généraliste issu d'un consensus international et multidisciplinaire. Il a pour objectif de fournir un socle commun d'éléments descriptifs suffisamment structuré pour permettre une interopérabilité minimale entre des systèmes conçus indépendamment les uns des autres. Le [Dublin Core](#) est un vocabulaire du web sémantique utilisé pour exprimer les données dans un modèle de type Resource Description Framework dans ses attributs (RDFa). Le Dublin Core définit un [ensemble d'items de métadonnées](#) obligatoires pour décrire les données :

1. Titre (métadonnée) (Title) : Nom donné à la ressource
2. Créateur (métadonnée) (Creator) : Nom de la personne, de l'organisation ou du service responsables de la création du contenu de la ressource
3. Sujet (métadonnée) ou mots clés (Subject) : Thème du contenu de la ressource (mots clés, expressions, codes de classification)
4. Description (métadonnée) (Description) : Présentation du contenu de la ressource (résumé, table des matières, représentation graphique du contenu, texte libre)
5. Éditeur (Publisher) : Nom de la personne, de l'organisation ou du service responsables de la mise à disposition ou de la diffusion de la ressource
6. Contributeur (Contributor) : Nom de la personne, de l'organisation ou du service responsables de contributions au contenu de la ressource
7. Date (métadonnée) (Date) : Date de création ou de mise à disposition de la ressource
8. Type (Type) : Nature ou genre de la ressource (catégories, fonctions, genres généraux, niveaux d'agrégation du contenu)
9. Format (Format) : Manifestation physique ou numérique de la ressource
10. Identifiant de la ressource (Identifier) : Référence univoque à la ressource dans un contexte donné (URI, ISBN)
11. Source (Source) : Référence à une ressource dont la ressource décrite est dérivée (URI)
12. Langue (métadonnée) (Language) : Langue du contenu intellectuel de la ressource
13. Relation (métadonnée) (Relation) : Référence à une ressource apparentée
14. Couverture (métadonnée) (Coverage) : Couverture spatio-temporelle de la ressource (domaine d'application)
15. Gestion de droits (métadonnée) (Rights) : Informations sur les droits associés à la ressource (IPR, copyright, etc.)

Pour la description des jeux de données géolocalisées, les [normes ISO 19115 et ISO 19139](#) sont des normes de référence dans le domaine des métadonnées pour l'information géographique. L'ISO 19115 fournit une structure permettant de décrire et de découvrir des données géospatiales, y compris le moment et l'endroit de leur localisation, une vue d'ensemble de leur contenu, de leurs propriétés, de leur qualité, de leur utilisation adéquate, du mécanisme de

distribution, des points de contact pour les demandes d'informations, etc. La norme ISO 19139 est l'implémentation XML de la norme ISO 19115. Elle définit le codage XML des métadonnées géographiques, une implémentation de schéma XML dérivée de la norme ISO 19115. La norme ISO 19139 est le modèle principal utilisé pour décrire des données dans le logiciel GeoNetwork et constituer ainsi un catalogue de données géospatialisées que l'on abordera dans le chapitre 7 « Publier » du présent guide.

3.1.3 Les protocoles standards en information Géographique

L'échange de données d'une plateforme à l'autre se fait au travers de protocoles informatiques. De ce fait, si l'on veut que les systèmes soient interopérables entre eux, il est nécessaire d'utiliser des protocoles ouverts et standards pour permettre l'interopérabilité. Dans le domaine environnemental, pour des données qui sont souvent géolocalisées par des coordonnées Latitude/Longitude, l'[Open Geospatial Consortium \(OGC\)](#), est un consortium international qui a pour objectif de développer et promouvoir des standards ouverts, les spécifications OpenGIS, afin de garantir l'interopérabilité des contenus, des services et des échanges dans les domaines de la géomatique et de l'information géographique.

Les standards OGC sont importants à connaître dans la mesure où ils définissent les protocoles et formats à suivre pour être interopérables. Ils ont été présentés par François André dans les réseaux DEVLOG et dans le réseau SIST de l'Institut National des Sciences de l'Univers (INSU). Pour ce dernier réseau, l'interopérabilité dans la gestion des données des Observatoires de l'INSU est un enjeu important.

Les Normes OGC (Open Geospatial Consortium)

François André, AerisSéminaire SIST15 - OSU Pytheas Marseille 2015

Parmi les standards de l'OGC les plus utilisés dans nos réseaux métiers chez les gestionnaires de données environnementales, on peut citer :

- **CS-W** - Catalog Service for the Web : ce protocole est destiné à diffuser des métadonnées ISO 19139 et permettre l'interrogation de catalogues de métadonnées. Une très bonne implémentation de ce protocole est réalisée dans le logiciel « [Geonetwork](#) » utilisé pour constituer des catalogues et des inventaires de jeux de données et les présenter sur le Web de manière interopérable. Ce logiciel est détaillé dans le chapitre 7 « Publier » du présent guide, dédié à la publication des jeux de données. Grâce à ce protocole, on peut constituer des réseaux de catalogues tels que demandés par la [Directive Européenne Inspire](#).
- **WMS** - **Web Map Service** est un protocole de communication standard qui permet de constituer des cartes de données géoréférencées à partir de différents serveurs de données cartographiques. .

Le réseau SIST a organisé deux actions de formation nationale (ANF) sur ces logiciels mettant en oeuvre les standards d'interopérabilité WMS, CSW et SOS qui permettent aux personnels d'améliorer la gestion et la diffusion de leurs données scientifiques d'observation en apprenant à installer, configurer et utiliser différents outils logiciels, choisis pour leur aptitude à répondre de manière standardisée à ces problématiques.

« Gestion des données d'observation : les outils informatiques pour la valorisation »

ANF SIST17, Fréjus - ANF SIST18, Autrans

De nombreux instituts et auteurs, gestionnaires de données suivent ces standards OGC :

Sylvain Grelet communique par exemple le retour d'expérience sur l'utilisation et le déploiement des standards d'interopérabilité au BRGM :

De la définition au déploiement de standards d'interopérabilité : retour d'expérience de la Direction des Systèmes d'Information (DSI) du BRGM

Grellet Sylvain, Stéphane Loigerot, BRGMSéminaire SIST15, Marseille

Véronique Chaffard nous présente la mise en oeuvre des standards de l'OGC dans le projet AMMA-CATCH :

Portail Web d'accès aux données de l'observatoire AMMA-CATCH et mise en oeuvre des standards d'échange des données OGC

Véronique Chaffard, IRDSéminaire SIST15, OSU Pytheas Marseille

3.2 Les systèmes d'acquisition : maîtriser l'acquisition et la collecte des données

Il est important que le processus de collecte des données soit clairement défini et validé. Par exemple, il conviendra de s'assurer que les systèmes d'acquisition sont bien étalonnés. Par ailleurs, l'ensemble des données produites doit être parfaitement répertorié et enregistré. Nous disposons pour ce faire d'un certain nombre de supports tels que les cahiers de laboratoires, les carnets de terrain. . .

3.2.1 La métrologie des équipements

Par nature, la recherche n'est pas un processus répétitif, elle est pleine d'aléas et d'incertitudes contrairement à un processus industriel. La confiance dans la qualité d'une recherche consiste à établir et vérifier que les différentes étapes d'une étude peuvent être répétées en obtenant le même résultat par différents chercheurs à des moments différents. Il est donc essentiel de s'assurer que l'ensemble des activités soient tracées et maîtrisées ; cela est une nécessité pour toute la chaîne fonctionnelle d'une analyse (des pipettes, balances jusqu'aux équipements d'analyse).

Confirmation métrologique des équipements

Virginie JAN LOGASSI, DAPEQ LUEANF Outils qualité, réseau QeR, 2019

De nombreux laboratoires et plateformes de tests du CNRS sont équipés de salles propres, dans des domaines variés tels que la micro et nanotechnologie, la géochimie, l'optique, la médecine, le spatial. . . En débutant par un point sur l'état de l'art (définition, réglementation, documentation. . .) de ces deux aspects, l'objectif principal de la journée thématique est de faire bénéficier de retours d'expériences sur les bonnes pratiques déjà éprouvées et sur les écueils à éviter afin de répondre, entre autres, aux questions suivantes :

- Quand a-t-on besoin de travailler en salles propres ?
- Quelles réglementations régissent l'installation, la maintenance et le contrôle des salles propres ?
- Comment préparer l'installation dans nos locaux ? A quoi doit-on penser ?
- Quelles sont les solutions techniques les mieux adaptées à notre besoin ?
- Quels sont les critères de surveillance et systèmes de contrôle des installations ?
- Comment doit-on travailler en salles propres ? Quelles sont les bonnes pratiques de gestion d'une salle propre ?

Les salles propres de l'installation à l'utilisation, de la théorie à la pratique - Usages et retours d'expériences

Journée thématique, réseau QeR, 2017

3.2.2 Les capteurs

Diverses communautés scientifiques sont intéressées par les problématiques inhérentes aux systèmes d'acquisitions et aux instruments associés. Différents aspects de collecte de données existent, qu'ils proviennent d'un équipement, d'un capteur automatisé, d'un modèle numérique ou qu'ils soient obtenus par un personnel de terrain, par une enquête, au moyen d'interfaces. Dès lors, il convient d'élaborer des méthodologies de collecte, de se documenter sur les choix des référentiels de métadonnées et des thésaurus de vocabulaire, mais également de développer les procédures d'intégration des données dans les bases.

Pour la thématique « Ocean-Atmosphere » cette problématique occupe une place importante, à tel point que, depuis plusieurs décennies, Météo-France et l'INSU depuis 1966, l'IFREMER depuis 2002, l'IRD et le CNES depuis 2004, le Service hydrographique et océanographique de la Marine (SHOM) depuis 2005, organisent un atelier dédié aux rencontres portant sur l'expérimentation et l'instrumentation. Cet [Atelier Expérimentation et Instrumentation \(AEI\)](#) permet de réunir la communauté scientifique spécialisée dans la recherche instrumentale et de traiter divers thèmes d'actualité lors de ses [différentes éditions](#). L'AEI traite de manière privilégiée les aspects de mesure et de méthodologie, sans exclure pour autant l'exploitation scientifique des résultats. Il a lieu alternativement à Paris, Toulon, Lille et Brest, généralement en début d'année. L'AEI permet aux équipes de recherche d'exposer leurs résultats dans un colloque francophone. C'est un lieu de rencontre pour les participants, issus des différents organismes et groupes industriels, afin de favoriser les synergies et coopérations.

Pour la gestion des capteurs, l'[OGC \(Open Geospatial Consortium\)](#) cité précédemment, publie un standard d'interopérabilité, [Sensor Web Enablement \(SWE\)](#), qui permet de présenter des données de capteurs de manière standardisée et interopérable. Ce protocole et les logiciels qui les implémentent sont bien adaptés à la description des capteurs et à la gestion des séries temporelles.

Le protocole « [SOS](#) » ([Sensor observation service](#)) de l'[OGC](#) permet de présenter de manière standardisée les données issues de capteurs de terrain de manière interopérable. Ce standard définit une interface de service Web qui permet d'interroger les observations, les métadonnées des capteurs, ainsi que les représentations des caractéristiques observées. En outre, cette norme définit les moyens d'enregistrer de nouveaux capteurs et de supprimer les capteurs existants. Elle définit également les opérations permettant d'insérer de nouvelles observations de capteurs.

Sensor Web Enablement Standards & Technology

Christoph Stasch, Simon Jirka, [52North Séminaire SIST15, Marseille](#)

Actuellement on trouve deux implémentations intéressantes du protocole SOS dans la gestion des données de capteurs environnementaux. Il s'agit de :

- [52North](#), logiciel de la société éponyme, est une application qui fournit une interface web interopérable pour l'insertion et l'interrogation des données et des descriptions des capteurs. Il regroupe les observations provenant de capteurs in-situ en direct ainsi que des ensembles de données historiques (données de séries chronologiques).
- [istSOS](#) est une implémentation de serveur OGC SOS écrite en Python. [istSOS](#) permet de gérer et d'envoyer des observations provenant de capteurs de surveillance selon la norme [Sensor Observation Service](#). Le projet fournit également une interface utilisateur graphique qui permet de faciliter les opérations quotidiennes et une [api RESTFull Web](#) pour automatiser les procédures d'administration.

[istSOS](#) est un logiciel libre qui fonctionne sur toutes les principales plates-formes (Windows, Linux, Mac OS X), même s'il n'a été utilisé en production que dans l'environnement Linux.

Présentation du logiciel istSOS

Massimiliano Canata [Séminaire SIST15, Marseille](#)

Ces 2 logiciels ont été présentés par Christoph Stasch, et Massimiliano Canata lors du séminaire du réseau [SIST](#) en 2015 à l'[OSU Pytheas Marseille](#).

Stephane Debaré présente l'utilisation d'[istSOS](#) dans la gestion de mesures altimétriques radars :

Mise en accord de mesures altimétriques radars avec le standard de l'OGC - SOS

Stéphane Debard IRDSéminaire SIST19 OMP Toulouse

3.2.3 Les chaînes de collecte

Les gestionnaires de données environnementales mettent en place des chaînes de collecte de données provenant de capteurs de terrains ou de modèles numériques. Ils se préoccupent de l'utilisation de normes interopérables dans les protocoles d'échange et dans les formats de données.

Régis Hocdé et ses collègues nous présentent un retour d'expérience sur le réseau de suivi de température des eaux côtières dans la région du Pacifique Sud et Sud-Ouest :

Retour d'expérience sur le système d'information dédié capteurs et reconstitution de séries temporelles de ReefTEMP

Sylvie Fiat, Régis Hocdé, Institut de Recherche pour le DéveloppementSéminaire SIST15, Marseille

Réseau d'observation du Pacifique Sud 'ReefTEMPS' : évolutions fonctionnelles et optimisation d'un système d'information dédié capteurs et reconstitution de séries temporelles

Régis Hocdé, Sylvie Fiat, Guillaume Brissebrat, Bernard Pelletier, Institut de Recherche pour le DéveloppementSéminaire SIST16, OSU OREME, Montpellier

Alban Thomas nous présente la technologie utilisée à base de Raspberry et de développement en Python, dans la constitution d'un réseau de stations météorologiques de la région rennaise.

Collecte de mesures météorologiques à l'aide d'un système autonome : exemple de la métropole rennaise (Zone Atelier Armorique)

Alban Thomas - Hervé Quénol UMR LETG RennesSéminaire SIST15, OSU Pytheas Marseille

3.2.4 Web scraping ou grattage Web : collecte automatique et analyse de données

« Le Web scraping est une technique permettant de convertir des données présentes dans un format non structuré (balises HTML) sur le Web en un format structuré facilement utilisable. Les exemples peuvent aller du texte sur Wikipedia, à des images sur Flickr en passant par les commentaires sur TripAdvisor, les articles d'actualité ou de chercheurs ou n'importe quelle page web présente sur Internet » [Introduction au Webscraping](#).

Depuis l'explosion quantitative des données numériques, il est devenu extrêmement intéressant d'apprendre à recueillir, comprendre et exploiter les informations issues du web. On constate ces dernières années, dans le domaine des sciences sociales, l'intérêt croissant des chercheurs ou ingénieurs pour l'utilisation de nouvelles techniques de collecte et de traitement automatisé des données et en particulier des données massives. Chaque utilisateur en fonction de son profil et de ses compétences peut choisir une technologie partant de simples outils comme les aspirateurs de site qui permettent de réaliser des opérations basiques de grattage (scraping) jusqu'à l'utilisation de langages plus performants comme R ou Python pour des utilisateurs plus avancés.

Au-delà des fonctionnalités de grattage web, la présentation « [Analyse de données avec R](#) » proposée par Hugues Pécout (CNRS) donne un exemple de l'analyse de donnée avec le logiciel R. En plus d'une présentation du logiciel R et de RStudio, elle contextualise R dans le paysage de l'analyse de données en le comparant à des logiciels propriétaires

existants sur le marché ainsi qu'au langage Python. En Python, il faut utiliser le package BeautifulSoup, qui est très populaire [Webscraping avec Python](#).

Ces outils sont depuis quelques années en plein essor car ils permettent d'automatiser la constitution des bases de données, de collecter des sommes de données importantes, inaccessibles il y a de cela quelques années comme les données de réseaux sociaux, de compiler des données pour créer ses propres indicateurs (impossible avec des techniques de collecte classiques) ou encore de nettoyer, structurer des données déjà existantes. . . Ces modes de collecte automatisés renvoient aussi aux notions d'exploration de données (Data Crawling) et de récolte de données (Data Harvesting).

Dans la pratique, des questions juridiques peuvent se poser au regard de l'exploitation des données récoltées en masse par ces moyens car ces données sont susceptibles d'être des données personnelles ou protégées par la propriété intellectuelle.

3.2.5 Les cahiers de laboratoire

L'ensemble des données produites par la recherche doit être répertorié et enregistré dans l'objectif d'une réutilisation potentielle. Nous disposons pour ce faire d'un certain nombre de supports comme les cahiers de laboratoire. Le cahier de laboratoire est un outil non obligatoire, mais fortement recommandé pour toute structure générant des données donnant lieu à des connaissances diffusables et valorisables. Il constitue un véritable outil scientifique et ce, dès le commencement d'un projet. Les cahiers de laboratoire répondent également aux obligations légales et contractuelles, en apportant la preuve de l'invention et de ses inventeurs. Les plaquettes du réseau CURIE « [Le cahier de laboratoire national : Pourquoi l'utiliser ?](#) » et « [Le cahier de laboratoire national : Comment l'utiliser ?](#) » présentent des recommandations sur la bonne gestion de ce dernier.

Alain Rivet positionne le cahier de laboratoire comme un outil de gestion des données de la recherche :

Cahier de laboratoire et gestion des données de la recherche

Alain Rivet, CERMAVAtelier Dialog'IST « [Rendre FAIR les données, mais quelles données préserver ?](#) », réseau Renatis, 2020

Les apports du numérique sont multiples en améliorant la traçabilité des recherches, la lutte contre la fraude et la gestion des données. Les cahiers de laboratoire électroniques présentent plusieurs ainsi avantages par rapport à leur version papier :

- le partage de l'information avec un rattachement des données brutes ;
- une recherche d'informations facilitée ;
- une datation assurée des expériences par l'horodatage.

Le site [dataacc.org](#) consacre la mise en œuvre d'un service d'accompagnement sur la gestion des données en physique et en chimie, dans le cadre d'un projet CollEx-Persée. Le site fournit des contenus nourris sur les cahiers de laboratoire électroniques, issus d'une expérimentation menée avec des chimistes de Lyon 1 et de Grenoble, assortis de [bonnes pratiques](#) sur leur utilisation.

Diverses expérimentations au sein de structures de recherche ont été réalisées :

Les cahiers de laboratoire électroniques : atelier elabFTW

Alain Rivet, Henri Valeins, CNRSEcole QUARES, Montpellier, 2020

Utilisation du cahier de laboratoire électronique BIOVIA au sein de l'Institut de Biologie Structurale

Cédric Laguri, IBSANF « [Traçabilité des activités de recherche et gestion des connaissances](#) », Réseau Qualité en Recherche, Grenoble, 16-18 octobre 2017

L'INSERM s'est fortement intéressé à la version numérique des cahiers de laboratoires, comme une réplique du cahier papier. L'INSERM pense que si la version électronique reste une solution d'enregistrement au quotidien des expériences scientifiques, c'est désormais devenu un outil différent, fortement axé sur la qualité, la gestion de la connaissance, la gestion de projets et le travail collaboratif. Paul-Guy Dupré et ses collaborateurs présentent les cahiers de laboratoires qui ont été mis en place à l'INSERM :

Expérimentation du cahier de laboratoire électronique à l'Inserm

Paul-Guy Dupré, INSERMANF « Traçabilité des activités de recherche et gestion des connaissances », Réseau QeR, Grenoble, 2017

Expérimentation du cahier de laboratoire électronique à l'Inserm : les apports de l'électronique au cahier de laboratoire

Paul-Guy Dupré, Fanny Brizzi Inserm, DSIJRES2017

Déploiement du cahier de laboratoire électronique à l'INSERM et nouvelles perspectives

PaulGuy Dupré Inserm & Claudia Gallina-Muller - Inserm DSIJRES2019

3.2.6 Les tablettes et carnets de terrain

Les données et documents produits directement sur le terrain témoignent de l'activité de recherche dans diverses disciplines, notamment en sciences humaines et sociales, en sciences de la terre... Il s'agit aussi bien de carnets issus d'entretiens de sociologues, d'ethnologues, de carnets de prélèvements en géochimie, géologie que de carnets de fouilles en archéologie, de notes, de photographies prises sur le terrain, etc. De plus, certaines données peuvent se révéler d'une valeur inestimable, qu'il s'agisse de données fortement temporelles (images satellites de la banquise, données sur les glaciers alpins) ou de données provenant de sites aujourd'hui endommagés ou détruits (Notre Dame de Paris, cité antique de Palmyre, etc). Il est de ce fait essentiel que ces données soient répertoriées et archivées.

L'utilisation de carnets de terrain électroniques que sont les tablettes permet de profiter des avantages d'appareils nomades pour faciliter la saisie des observations que l'on fait sur le terrain, en milieu naturel. L'utilisation de cet outil « nomade » va permettre :

- d'améliorer la qualité des données collectées ;
- de pouvoir utiliser les données plus rapidement ;
- de réduire le coût (temps de ressaisie).

Cependant, ces nouvelles technologies très « ludiques » et « faciles » d'utilisation, nécessitent une réflexion importante pour définir de façon précise son besoin afin de ne pas être pénalisé sur le terrain. Elles nécessitent aussi une adaptation technologique pour permettre un stockage efficient et pérenne en bases de données.

Au niveau logiciel, cinq stratégies sont possibles pour développer des carnets de terrain électroniques :

1. utiliser une application nomade existante
2. utiliser une application web existante
3. développer une application nomade spécifique avec un langage de programmation
4. développer une application nomade en utilisant une boîte à outils de génération de carnets de terrain
5. développer une application nomade en adaptant des logiciels existants (par exemple QGIS, Lizmap)

Deux solutions ont été étudiées au Centre d'Ecologie Fonctionnelle et Evolutive (CEFE) : le développement d'une application nomade basée sur le système d'information géographique, libre, multiplate-forme, publié sous licence GPL QGIS ainsi qu'une application nomade utilisant une boîte à outils de génération de carnets de terrain électronique [Open Data Kit](#).

Dans la présentation détaillant la solution basée sur QGIS, l'auteur détaille l'étude et le développement de l'application interopérable avec le système d'information du laboratoire CEFE et qui permet aux intervenants sur le terrain de collecter les données :

Carnet de terrain électronique, Retour d'expérience sur la création d'une boîte à outils

Marie-Claude Quidoz, CEFE 15èmes Rencontres Mondiales du Logiciel Libre, Montpellier, 2014

La solution basée sur ODK a servi de fil rouge à l'ANF « [Interfacer les outils mobiles avec son système d'information](#) » en 2019, car la solution ODK permet de couvrir les étapes allant de la création du formulaire à la sécurisation en bases de données.

Des applicatifs « clef en main » ont été développés à partir du moteur ODK. Le plus connu est sans doute [KoboToolbox](#), qui, aux fonctionnalités de base, a ajouté quelques fonctionnalités supplémentaires telles que le FormBuilder et la bibliothèque de questions.

Collecte de données terrain avec un smartphone : Prise en main de Kobotoolbox et de Kobocollect

Akaouette, Ata Franck FOSS4G-fr, Marne-la-vallée, 2018

Pierre-Yves Arnould nous présente sa solution à base de ODK pour faciliter la saisie, rendre autonome les chercheurs dans leur saisie, uniformiser la structure des fichiers, génération d'étiquettes pour les échantillons, et sauvegarder automatiquement sur un micro-serveur sur le terrain puis sur le SI OTELo

Retour terrain : la délicate question de l'intégration des données

Pierre-Yves Arnould, OTELoANF « [Interfacer les outils mobiles avec son système d'information](#) », réseau RBDD, 2019, Sète.

De nombreuses autres solutions sont aussi envisageables, nous invitons le lecteur à consulter les ateliers et séminaires suivants pour en découvrir leurs avantages et inconvénients :

Atelier « Carnets de terrain électroniques »

Réseau Zones Ateliers, Montpellier, 2018

Séminaire « [Système d'information embarqué, cahier/carnet de terrain et de laboratoire électronique : quelles interactions avec les bases de données ?](#) » Réseau rBDD, Paris, 2016

Il est à noter que la collecte sur le terrain nécessite de s'équiper d'un matériel apte à être utilisé sur des terrains parfois hostiles. Le choix de l'équipement conditionne aussi le choix de la solution logicielle comme le montre Marie-Claude Quidoz lors de cette présentation :

Carnet de terrain électronique

Vidéo : Marie-Claude Quidoz, CEFESéminaire « [les technologies mobiles : retours d'expériences et perspectives](#) », Réseau ResInfo, Paris, 2016

3.2.7 La gestion des collections

Collec-Science est un logiciel web qui a été créé pour suivre les échantillons collectés lors des campagnes d'acquisition, et permet de répondre, entre autres, à ces questions :

- où est stocké l'échantillon ?
- d'où vient-il, quelle est sa généalogie (protocole de collecte, métadonnées associées à l'échantillon et ceux de ces ancêtres) ?
- quelles transformations ou opérations a-t-il subies ?
- sous quelle forme est-il conservé, existe-t-il un risque à le manipuler ?

Fruit d'une collaboration initiale entre l'Irstea (centre de Bordeaux), le laboratoire Epec à Bordeaux, le LIENSs à La Rochelle, il a été enrichi avec la participation de nombreux autres laboratoires, dont les laboratoires Chrono-environnement à Besançon, Edytem à l'Université Savoie - Mont Blanc, etc. Il a été choisi par le Réseau des Zones Ateliers pour assurer le suivi des échantillons.

Stockez et retrouvez vos échantillons avec Collec-Science

Marie-Claude Quidoz

3.3 Environnements de stockage - Sauvegarder les données

Dès la phase de collecte, il convient de se préoccuper des aspects de stockage et de sauvegarde qui seront plus largement abordés dans la phase 6 du cycle de vie des données. En effet, dès le début d'un projet, il est nécessaire, d'une part, d'estimer le stockage nécessaire à la collecte de données et d'autre part, de mettre en place les moyens de sauvegarde des données récoltées. La duplication des données par stockage redondant sur des supports différents de ceux de l'équipement utilisé (poste de travail fixe, mobile, serveur, ...) est un des principes de base d'une bonne conservation. Il convient de préférer un archivage centralisé conformément à la règle du 3-2-1 généralement recommandée (3 copies sur 2 supports différents dont 1 sur un lieu déporté). À cet effet, il conviendra de travailler en amont avec une équipe informatique afin que les dispositifs de stockage soient disponibles.

Rappels théoriques concernant les architectures de stockage traditionnel

Sylvain MaurinANF « Stockage Distribué », 2016

Outils algorithmiques et logiciels pour le stockage distribué

Benoit ParreinANF « Des données au BigData : exploitez le stockage distribué ! », 2016

Divers outils de sauvegarde des données sont fréquemment utilisés dans les milieux informatiques comme [backuppcc](#), [bacula](#), [rdiff-backup](#)

Un nouveau paradigme dans la sauvegarde consiste à introduire et utiliser des fonctionnalités de *déduplication*. Cette technologie consistant à réduire les volumes sauvegardés et les durées de sauvegarde en découpant les gros fichiers en fragments (blocs) et en ne sauvegardant qu'une seule fois les fragments identiques.

Un retour d'expérience sur le [logiciel borgbackup](#) donne des résultats intéressants et prend tout son sens quand on a beaucoup de fichiers volumineux peu différents.

Sauvegardes déduplicées avec BorgBackup : retour d'expérience

Maurice Libes - Didier Mallarino, OSU PytheasJRES 2017, Nantes

Respecter le RGPD!

Enfin n'oublions pas que, dès lors que l'on collecte des données personnelles (données permettant l'identification directe ou indirecte d'une personne), il est important de respecter des principes essentiels sur la durée de conservation des données, le droit à l'information et l'obligation de sécuriser les données. Il ne faut pas hésiter à se rapprocher du correspondant du Délégué à la protection des données (DPD) de votre délégation (pour le CNRS) ou du Délégué à la protection des données de votre établissement.

Cette phase du cycle de vie des données correspond au prétraitement des données brutes issues des acquisitions et des collectes. Il s'agit souvent de regrouper, choisir, qualifier les données pertinentes parmi celles qui ont été collectées, puis les reformater dans des formats standards interopérables, et les préparer en vue de leur analyse ultérieure.

Cette partie est donc structurée en différentes sections décrivant cette préparation des données :

- Préparer les fichiers de données, en vue de leur analyse, en utilisant des formats interopérables.
- Utiliser des infrastructures logicielles « framework » d'intégration de données, lorsqu'elles sont hétérogènes.
- Mettre en place et utiliser des plateformes de gestion de données locales, en vue de leur analyse.
- Vérifier et s'assurer de la qualité des données.

4.1 Préparer les fichiers de données en vue de leur analyse

Bien souvent, les données « brutes » sont issues de capteurs ou divers instruments de collecte sur le terrain. Ils se présentent fréquemment sous la forme de fichiers dans des formats propriétaires, peu exploitables et peu interopérables directement tels quels.

Dans une optique de gestion FAIR, il est donc important de se préoccuper du format des données afin de les rendre « ouverts » et interopérables. La notion de format « ouvert » est importante pour que les données puissent être partagées, interopérables et préservées sur le long terme. A cet effet, le site Doranum propose une [introduction à la définition de formats ouverts ou fermés](#).

De plus, si l'objectif est le traitement massif des données, il est important de choisir des formats capables de supporter des entrées / sorties intensives sur des infrastructures de calcul.

4.1.1 Utiliser des formats standards

Parmi les premiers traitements opérés sur des données brutes provenant du terrain, les données issues de capteurs environnementaux sont souvent illisibles et peu exploitables par un être humain. Il convient alors de traiter les fichiers bruts de manière à en extraire les données utiles, et de les réécrire dans des formats standards utilisables par un grand nombre de logiciels, et une communauté d'utilisateurs.

Chaque discipline utilise, voire définit un certain nombre de formats standards, et il est bon de les connaître et de s'y référer.

On ne pourra pas tous les citer, mais à titre d'exemple dans les domaines Océan, Atmosphère par exemple,

- **Le format NetCDF** est un format ouvert, autodocumenté et très utilisé en particulier dans les communautés sciences de l'environnement. Il est très bien adapté et utilisé, par exemple pour représenter et formater des données qui sont des profils verticaux, des séries temporelles, des trajectoires, ou encore des surfaces maillées en 2D. Ce format est dit « auto-descriptif » en ce sens qu'il permet de ne pas avoir besoin d'un fichier de description complémentaire. Les métadonnées sont en effet insérées dans l'entête du fichier, avec les données elles-mêmes. On peut ainsi décrire de manière assez précise les données du fichier, par exemple en insérant les unités de mesure des paramètres mesurés, la licence de diffusion, les propriétaires, etc., ainsi que l'organisation des données.

Toutefois dans son format originel NetCDF n'a pas imposé de directives particulières pour inscrire les métadonnées dans l'entête du fichier. De ce fait, il était possible d'inscrire n'importe quel libellé de variables, unités, etc. Une standardisation a été nécessaire pour obtenir des fichiers compréhensibles et interopérables. C'est le but de la [convention CF \(climate forecast\)](#) qui fournit une [table de standardisation des variables et unités de mesures](#) à inscrire dans l'entête d'un fichier NetCDF.

Ce format standard, la convention « CF », et l'interface de programmation (API) en Python pour créer des fichiers NetCDF par programme ont été présentés au [séminaire SIST19 à l'OMP](#) de Toulouse, par Joël Sudre, Maurice Libes et Didier Mallarino :

Présentation du format NetCDF

Joël Sudre, LEGOS [Séminaire SIST19 Toulouse](#)

La convention CF (climate forecast) pour les fichiers NetCDF

Joël Sudre, LEGOS et Maurice Libes, Institut Pytheas [Séminaire SIST19 Toulouse](#)

Utilisation de l'API de programmation Python pour NetCDF

Maurice Libes, Didier Mallarino, Institut Phyteas [Séminaire SIST19 Toulouse](#)

- **Le format ODV** (ocean data view) est également un format standard ouvert intéressant. C'est un format de type « tableur », ensemble de lignes comportant un nombre fixe de colonnes qui se rapproche d'un format CSV, composé de colonnes de données séparées par des virgules (ou tout autre séparateur), à cette différence près que le format ODV permet l'insertion d'un entête assez riche permettant de placer des métadonnées en début de fichier. On trouvera un exemple sur le [Portail des données marines](#).

Le format de données ODV permet un stockage dense et un accès très rapide aux données. De grandes collections de données comprenant des millions de stations peuvent être facilement entretenues et explorées sur des ordinateurs de bureau.

Un explorateur et extracteur de données webODV est disponible sur le portail [EMODnet Chemistry](#). L'outil webODV Data Explorer and Extractor, développé à l'Institut Alfred Wegener en Allemagne, permet aux utilisateurs d'explorer, de visualiser et d'extraire des sous-ensembles de données validées simplement en utilisant leur navigateur web.

Les formats NetCDF et ODV sont les formats recommandés et utilisés par le [pôle de données Odatis](#) et par le projet européen [Seadatanet](#).

— **Le format HDF5**

Le format **HDF5** (Hierarchical Data Format, version 5) est un format de fichier de type conteneur, c'est-à-dire assimilable à une arborescence de dossiers / fichiers contenus dans un même fichier.

C'est un format très utilisé lorsqu'on veut traiter ou simuler des données grâce au calcul intensif, car il offre des possibilités de compression et d'écriture/lecture parallèles très efficaces.

Des supports de formation sur ce format sont de ce fait disponibles via les infrastructures et réseaux en lien avec le calcul intensif :

Formations PRACE

HDF5 : theory & practice 1 et 2

Prace Advanced Training Centers, [Course: Parallel I/O and management of large scientific data, 2014](#)

4.2 Organiser les données

4.2.1 Développer les procédures d'intégration des données dans les bases de données

Les nouveaux mécanismes de collecte de données ont souvent simplifiés la mise en base de données comme c'est le cas avec la boîte à outils ODK (cf partie [Collecter](#)) qui envoie directement les données collectées sur tablette dans un schéma d'une base de données PostgreSQL. Mais, pour sécuriser les données, elles doivent être ensuite transférées dans la base de données métier. Cette opération est souvent réalisée à l'aide de déclencheur comme on le voit dans la présentation suivante

Intégrer les données dans sa base métier

Marie-Claude Quidoz, CEFEANF « Interfacer les outils mobiles avec son système d'information », Réseau rBDD, Sète, 2019

4.2.2 Utiliser un cadre d'applications d'agrégation de données

Lorsque les données à traiter sont hétérogènes et que les technologies qui permettent de les fournir sont également différentes, une solution est d'utiliser un « framework » d'agrégation de données. Un « framework » est un cadre d'applications d'agrégation de données, autrement dit un outil qui va permettre de traiter des données de formats différents de façon transparente pour l'utilisateur final.

Le logiciel « [Lavoisier](#) » développé au Centre de Calcul de l'IN2P3 (CC-IN2P3), permet de récupérer, transformer, fusionner, et requêter des données de sources différentes. Il est utilisé dans plusieurs contextes pour fournir une vue unifiée des données collectées à partir de multiples sources hétérogènes

Lavoisier : un cadre d'applications d'agrégation de données, vidéo de la présentation

Cyril L'Orphelin, Sylvain Reynaud, CC-IN2P3, CNRSJCAD 2018, Lyon.

D'autres outils logiciels existent, permettant l'intégration de données. Dans la catégorie des logiciels « ETL » (Extract, Transform, Load, le logiciel « [Talend Open Studio](#) » par exemple, a été abordé lors d'une session de formation du réseau RBDD :

« Utilisation et maîtrise d'un ETL : intégrations de données avec Talend Open Studio »

Eric Quinton Réseau RBDD, 2017. Paris.

Ce logiciel « Talend » a été également utilisé par Soumaya Lahbib pour traiter les fichiers de données issues des capteurs du [projet EMSO Ligure-ouest](#) et les transformer en fichiers CSV utilisables facilement.

Gestion des données du projet EMSO avec Talend et ERDDAP

Maurice Libes, Soumaya Lahbib Séminaire SIST18 OVSQ, Guyancourt.

4.2.3 Déposer et structurer dans des plateformes de gestion de données locales

Après la phase de collecte de données que nous avons vue dans l'étape précédente du cycle de vie des données, il est nécessaire de se préoccuper du dépôt, de la facilité d'accès et de la réutilisation des données localement dans une unité de recherche.

Un certain nombre de logiciels font office de plateforme d'accès et de gestion des données. Ils permettent de présenter les données avec leurs métadonnées, de fournir des interfaces de recherche, de géolocaliser les données, et parfois de visualisation des données avec des graphes. Cette organisation des données facilite grandement leur analyse ultérieure.

Des logiciels sont particulièrement adaptés dans la diffusion et l'affichage des données scientifiques d'observation par le fait qu'ils utilisent les standards interopérables de l'Open Geospatial Consortium (OGC), comme le [protocole DAP](#) (Data Access Protocol)

- Les plateformes de dépôt et de diffusion de données comme [THREDDS](#) et [ERDDAP](#) sont intéressantes par le fait qu'elles mettent en oeuvre le protocole DAP, et sont des solutions très bien adaptées pour rendre les données FAIR et faciliter la diffusion des données.

La plateforme d'accès ERDDAP se présente comme étant un « accès facile aux données scientifiques » (« Easier access to scientific data ») et fournit un ensemble complet de fonctionnalités pour la gestion des jeux de données. Il permet :

- déposer des jeux de données dans différents formats interopérables
- de fournir un catalogue des jeux de données gérés par le serveur
- d'afficher les métadonnées inscrites dans les fichiers
- de lire et convertir des jeux de données dans de nombreux formats standards interopérables différents,
- d'interroger et filtrer les données au travers de formulaires,
- de créer des graphiques et des cartes simples pour visualiser le jeu de données analysé
- de normaliser le format des unités de temps présentes dans les fichiers.

Une des fonctionnalités intéressantes est qu'ERDDAP agrège automatiquement les données nouvelles répondant à un format donné, qui sont déposées dans un répertoire. Ainsi pour les séries temporelles cette fonctionnalité est intéressante puisqu'il suffit de déposer des fichiers dans un répertoire pour que la série soit automatiquement enrichie et mise à jour.

Dans le projet scientifique [EMSO](#), le logiciel ERDDAP permet de constituer un [réseau de serveurs](#) permettant de rassembler et fournir toutes les données d'un même projet sur plusieurs sites .

Utilisation de ERDDAP - Installation, configuration et mise à disposition de jeux de données

Maurice Libes, Didier Mallarino ANF SIST 17 Fréjus

Lorsque les données sont géoréférencées, on peut aussi déposer et faire gérer des données de terrain via des serveurs cartographiques comme :

- le serveur cartographique [Geoserver](#) permet d’afficher et d’échanger des données géospatiales sur le web selon les standards (WMS, WFS, ...) de l’OGC ;

Geoserver - Installation, configuration, affichage et diffusion de jeux de données géospatialisés

Juliette Fabre, Olivier LobryANF SIST 2018, Toulouse.

- l’application GeoCMS permettent la visualisation de données géospatiales sur le web et de mettre en place une Infrastructure de Données Géographique (IDG). on peut voir un exemple de l’interet de cette application sur un [portail comme celui de Indigeo](#)

TP GeoCMS - Installation, configuration, visualisation et interrogation de jeux de données géospatialisés

Mathias Rouan, Jonathan SchaefferANF SIST 2018, Toulouse.

Exemple de mise en oeuvre de plateformes de données

Des exemples d’utilisation des plateformes logicielles ERDDAP et THREDDS ont été présentés lors de différentes sessions des journées du réseau SIST :

G. Brissebrat nous montre comment sont diffusées des données maillées NetCDF du SEDOO avec la plateforme logicielle THREDDS. Les avantages évoqués sont nombreux :

- Consulter les métadonnées sans avoir à télécharger le jeu de données
- Accéder uniquement à une partie d’un jeu de données
- Télécharger un seul fichier même si les données originales sont réparties dans plusieurs fichiers
- Avoir le choix entre plusieurs moyens d’accéder aux données
- Accéder aux données dans des format compatibles avec les outils communs d’analyse ou de visualisation de données
- Offrir plusieurs formats et protocoles d’accès aux données-
- Fournir une prévisualisation des données
- Pouvoir moissonner des données d’un autre serveur THREDDS

Distribution et visualisation de données avec THREDDS, exemples d’utilisation au SEDOO

Guillaume Brissebrat, Service de données de l’OMPSéminaire SIST 2015 OSU Pytheas Marseille

Eccad, un exemple de mise en oeuvre de THREDDS

Sabine Darras, Observatoire Midi-PyrénéesSéminaire SIST 2019 OMP Toulouse

Dans cette présentation les auteurs montrent un workflow complexe depuis l’acquisition de données a 2500m de profondeur, jusqu’à l’affichage et la diffusion sur un serveur ERDDAP. Les données de capteurs sont traitées avec l’ETL Talend pour produire des fichiers CSV et NetCDF qui sont diffusées via la plateforme ERDDAP. Dans le cas de séries temporelles qui s’enrichissent quotidiennement, erddap permet d’aggréger automatiquement les données journalières qui sont déposées par programme dans un répertoire, sans intervention humaine.

Gestion des données du projet EMSO avec Talend et ERDDAP

Soumaya Lahbib, Maurice Libes, OSU PytheasSéminaire SIST 2018 OVSQ, Guyancourt.

La plateforme de gestion de données ERDDAP est utilisée dans le projet Européen EMSO et permet de constituer un réseau de serveurs qui regroupe les données d'un même projet avec des données issues de sites différents ¹.

Dans cette présentation, les auteurs avaient pour objectif de diffuser des données dans un environnement tropical et ont utilisé et comparés les plateformes logicielles THREDDS et ERDDAP

Copier les succès et rester simple (AMEO) : mise à disposition de sorties de modèles climatiques avec un NAS, THREDDS et ERDDAP.

Thierry Valéro, Institut de Recherche pour le Développement, Laboratoire d'Océanographie et du Climat Séminaire SIST 2016 OSU OREME Montpellier

Les présentations suivantes fournissent un certain nombre de connaissances sur l'utilisation d'infrastructure de données géographiques (IDS, IDG) et de différentes plateforme logicielle de gestion des données

Infrastructure de données spatiales et de traitements GEOSUD : des standards à la réalité

Jean-Christophe Desconnets, UMR Espace-Dev, IRDS Séminaire SIST 2016 OSU OREME Montpellier

Publication automatique de données et de métadonnées dans geOrchestra

Ernest Chiarello, Théoriser et modéliser pour aménager, MSHE Séminaire SIST 2018, Guyancourt.

Loïc Salaun nous montre un exemple de consultation des données à partir d'un visualiseur cartographique (visualiseur d'INDIGEO), utilisant les services web géographiques (WMS, WFS, WCS, CSW)

Mise en place d'une IDS pour le programme de recherche Réseau de Suivi et de Surveillance de l'Environnement.

Loïc Salaun, Observatoire des Sciences de l'Univers Nantes Atlantique Séminaire SIST 2016, Montpellier.

4.3 Mettre en place un contrôle qualité des données

Par nature, la recherche n'est pas répétitive, mais riche en incertitudes contrairement à un processus industriel. La confiance dans la qualité d'une recherche consiste donc à établir et vérifier que les différentes étapes d'une étude peuvent être répétées en obtenant le même résultat par des chercheurs différents à des moments différents. Ainsi, une donnée est fiable si, dans des conditions données, aucune déviation n'est constatée en fonction du temps, durant un laps de temps donné. Il est donc essentiel de s'assurer que l'ensemble des activités de recherche soit maîtrisé.

Le contrôle sur les équipements est le premier pas vers la traçabilité des données comme l'illustre l'exposé suivant :

En sciences environnementales, la qualification des données est importante pour estimer et fournir un degré de qualité de la donnée. On utilise pour cela des codes qui renseignent sur la qualité de la donnée : bonne, mauvaise, manquante, modifiée etc. . . Dans ce domaine bien souvent chacun utilise une codification personnelle, cependant une standardisation des codes qualité est bienvenue.

L'infrastructure de données européenne Seadatanet utilise par exemple une table « L20 » standardisant les codes qualité à placer sur les données

1. <http://erddap.emso.eu:8080/erddap/info/index.html?page=1&itemsPerPage=1000>

Traçabilité des données de la recherche. Confirmation métrologique des équipements

Virginie JAN LOGASSI, Université de Lorraine [Rencontres du réseau Qualité en Recherche](#), 2019, Nancy.

Le réseau rBDD a consacré un atelier à la qualité des données pour apporter des éclairages sur les questions suivantes :

- Quelles sont les différentes notions de qualité des données ?
- Comment contrôler la qualité des données dans la BDD : avant ou pendant l’insertion de données
- Faut-il automatiser le contrôle de la qualité dans les bases de données ?
- Quels sont les outils disponibles et comment les utiliser ?

Le programme de l’atelier s’appuie sur les travaux de [Laure Berti Equille](#) qui « classe les travaux autour de la problématique de la qualité des données selon quatre grands types d’approches complémentaires : prévenir / diagnostiquer / corriger / adapter ».

Dans la première partie de la présentation, après avoir explicité les notions autour de la qualité des données, Christine Plumejeaud nous donne de bonnes pratiques comme celle d’attribuer un code standard (suivant une norme choisie et citée) décrivant l’état de la valeur. Elle cite comme exemple le standard [SDMX](#), qui est une initiative internationale, utilisée entre autre par Eurostat et l’INSEE. Elle cite aussi les travaux faits par le Service d’Observation en Milieu Littoral [SOMLIT](#) qui a défini sa propre classification².

Sa présentation se poursuit sur l’utilisation de contraintes SQL pour éviter l’insertion en base de données de valeurs incohérentes ou impossibles. Ces contraintes sont la transcription des règles de gestion définies lors de la modélisation de la base de données. Une fois la structure de la base de données définie, il reste une étape, celle du nettoyage des données, à réaliser avant l’intégration des données en base. Le réseau rBDD conseille pour cela le logiciel [OpenRefine](#) très simple à prendre en main et très puissant.

Qualité des données

Christine Plumejeaud, LIENSs & Nadine Mandran, LIG ANF « Sciences des données : un nouveau challenge pour les métiers liés aux bases de données », réseau rBDD, Sète, 2018

Une présentation autour de l’outil OpenRefine de nettoyage et mise en forme des données.

Mathieu SABY, BU Université de Nice Sophia-Antipolis

Dans cette intervention, Christine Plumejeaud se place dans le cadre de l’utilisation d’outils nomades qui envoient directement les données collectées sur tablette dans une base de données. La problématique est sensiblement différente. Partant du principe que sur le terrain, il est communément recommandé de laisser la saisie la plus libre possible pour permettre une prise en compte des aléas plus faciles, la détection des choses non conformes aux règles métier est à traiter a posteriori.

Outils nomades : validation des données

Christine Plumejeaud-Perreau, CNRS, U.M.R 7266 LIENSs, la Rochelle ANF « Interfacer les outils mobiles avec son système d’information », réseau RBDD, 2019

Certains logiciels comme [ODV \(Ocean Data View\)](#) permettent de qualifier les données et d’attribuer un code qualité à des données après analyse par un expert du domaine. ODV est un format de fichiers, et un logiciel utilisés par le projet européen [SeadataNet](#).

Cependant peu de logiciels de traitement de données propose d’associer des codes qualités aux données, aussi on retrouve souvent de nombreuses méthodes et implémentations personnelles pour essayer de qualifier les données,

2. Codes qualité SOMLIT

illustrées par les exposés suivants donnés lors des journées de séminaires SIST (Séries Interopérables et Systèmes de Traitement) :

P. Téchiné présente les méthodes de suivi de la qualité de diverses mesures comme le niveau de la mer ou la salinité de surface (SSS : Sea Surface Salinity) dans différents projets. On peut constater la diversité des solutions mises en place.

Suivi de la qualité des mesures de réseaux d'observation océanographique

Philippe Téchiné, B. Buisson, L. Testut, T. Delcroix, G. Alory, Laboratoire d'études en Géophysique et océanographie spatiales [Séminaire SIST 2016 OSU OREME Montpellier](#)

Dans cette présentation Lynn Hazan décrit son processus d'attribution de code qualité. Comme nous l'avons indiqué précédemment dans la phase de traitement, les données sont obtenues en temps quasi-réel et sont transformées en données consolidées par un traitement qui permet d'en augmenter la précision et la confiance. Les étapes de consolidation incluent une expertise humaine avec une inspection visuelle afin de détecter des problèmes potentiels difficilement détectables automatiquement. L'outil ATCQc a été développé afin de permettre aux scientifiques de visualiser et qualifier rapidement leurs données issues des instruments de mesures du réseau.

ATCQc : Un outil pour le QA/QC de mesures atmosphériques du TGIR ICOS, vidéo

Lynn Hazan, Laboratoire des Sciences du Climat et de l'Environnement [Séminaire SIST 2018 OVSQ](#) , Guyancourt.

Dans cette présentation, les auteurs abordent la qualité des données sous l'angle utilisation de référentiels pour décrire finement les données et les rendre interopérables

La qualité des données à l'OSU OREME

Juliette Fabre, Olivier Lobry, Observatoire de REcherche Méditerranéen de l'Environnement [Séminaire SIST 2018 OVSQ](#), Guyancourt.

La qualité des données à l'OSU OREME

Juliette Fabre, Olivier Lobry, Observatoire de REcherche Méditerranéen de l'Environnement [Séminaire SIST 2018 OVSQ](#), Guyancourt.

Dans cette présentation, les auteurs proposent un développement graphique avec la librairie « [DyGraphs](#) » pour visualiser et valider des données de séries temporelles.

Outil web interactif de visualisation et validation de séries temporelles

Olivier Lobry, Juliette Fabre [Séminaire SIST 2015 OSU Pytheas Marseille](#).

Dans son projet A. Campos utilise un ensemble de scripts Python pour convertir les fichiers « xls » en fichier « ascii », puis effectue un nettoyage avec la commande « awk » de Unix. Enfin des scripts en langage R permet de faire des moyennes glissantes, des graphes et des exports des fichiers au format NetCDF.

Site Web de diffusion des données « Sahelian Dust Transect »

André CAMPOS, Laboratoire interuniversitaire des systèmes atmosphériques [SIST 2016 OSU OREME Montpellier](#)

Derrière le terme « analyser » s’entend l’extraction de l’information des données le plus souvent par l’utilisation de puissance de calcul. Cela recouvre de nombreux types de techniques (calcul intensif, traitement statistique, machine learning, visualisation . . .), et nécessite également des plateformes adaptées.

Cette étape du cycle de vie de nombreuses données impose que ces données soient exploitables, c’est-à-dire bien organisées, dans des formats adaptés à l’analyse envisagée, de façon à pouvoir leur appliquer des traitements automatisés.

Plusieurs évènements récurrents, annuels ou bisannuels, auxquels participent activement les réseaux métiers, comme les *JCAD (Journées Calcul et Données)*, les *JDEV (Journées du DEVeloppement logiciel)* par exemple, intègrent de nombreuses interventions sur ces différentes thématiques, allant de la description des plateformes aux outils disponibles, en passant par l’organisation des développements et la reproductibilité, détaillée dans la section *Reproductibilité* de ce guide. Ils incluent aussi très souvent des retours d’expérience particulièrement riches.

5.1 Plateformes de traitement de données

De nombreuses ressources sont disponibles, à différentes échelles, pour analyser et traiter des données. De façon générale, on distingue :

- Les ressources de type calcul intensif ou HPC (High Performance Computing) organisée à l’échelle européenne (EuroHPC ou Tier 0), nationale (GENCI et les centres nationaux ou Tier 1) et régionale (mésocentres ou Tier 2). Ces ressources sont adaptées aux simulations massives.
- Les ressources de type cloud (par exemple le cloud distribué de France Grilles : FG-Cloud). Ces ressources souples répondent aux besoins de calcul à la demande ou lorsque la maîtrise de l’ensemble du système est nécessaire.
- les ressources de type grille de calcul ou HTC (High Throughput Computing), par exemple l’infrastructure France Grilles ou le Centre de Calcul de l’IN2P3. Ces ressources sont utilisées pour faire du traitement massif de données.

Elles sont décrites dans la section *Infrastructures* de ce guide. Le choix du type d’infrastructures adapté au besoin n’est pas forcément trivial. Il est souvent plus pertinent de s’adresser à des spécialistes qui sauront vous orienter. En général, les mésocentres de calcul, grâce à leur proximité et à leur connaissance du domaine, sont de bons conseils. Une liste est disponible sur le [site du réseau Calcul](#).

5.2 Outils pour l'analyse des données

Il existe de très nombreux outils permettant d'analyser ses données, allant du langage de programmation au workflow de traitement en passant par les logiciels de visualisation.

5.2.1 Langages de programmation

Certains langages de programmation sont plus particulièrement utilisés pour l'analyse de données. En dehors du langage R spécifique aux statistiques et à la science des données, l'écosystème s'enrichit très rapidement :

- [Python](#) devient de plus en plus utilisé en science des données. Une introduction sur le sujet a été réalisée en décembre 2017 par Francis Wolinski (Société Yotta Conseil) dans le cadre d'une journée organisée par le réseau Calcul.

Présentation et illustration de l'écosystème Python pour la data science

Francis Wolinski (Société Yotta Conseil) Journée Python et Data Science IRMAR Rennes - 2017

- [Julia](#) est un des langages qui prend de l'importance sur ce sujet. Plusieurs présentations, lors d'une [journée d'introduction au langage](#) organisée par le réseau Calcul en janvier 2019, apportent un éclairage intéressant, en particulier le cadre des algorithmes Map/Reduce, ainsi que les performances du langage sous forme de benchmarks.

Map/Reduce operations for scientific computing in Julia

Xavier Vasseur (ISAE-SUPAERO) [Journée Julia - Lyon 2019](#)

Julia : benchmark et bonnes pratiques

Benoît Fabrèges (Institut Camille Jordan, Lyon) [Journée Julia - Lyon 2019](#)

Des retours d'expérience illustrent l'utilisation de ces outils.

Concernant les outils Python, l'utilisation de Dask à la place de job array a été présentée lors des JCAD 2019 par Guillaume Eynard-Bontemps, CNES. Dask est une bibliothèque parallèle Python qui facilite l'exécution massive de calculs sur des données distribuées.

Simulation paramétrique : Passez d'un job array à Dask

Guillaume Eynard-Bontemps (CNES), [JCAD 2019, Toulouse](#).

5.2.2 Approches méthodologiques

L'analyse des données ne concerne pas uniquement les modèles statistiques. De nombreux domaines appliqués reposent sur l'analyse de données géométriques : médecine, neurosciences, sismique, météorologie, vision par ordinateur, apprentissage statistique. Cette variété d'applications se retrouve dans la forme, la qualité et la sémantique des données ainsi que dans la nature des problèmes mathématiques qu'elles posent. Une école thématique a été consacrée à ce sujet en 2018, à destination des non spécialistes. Elle l'a abordé sous plusieurs angles :

- Analyse topologique de données,
- Anatomie computationnelle,
- Méthode d'évolution de front et fast marching,
- Méthodes variationnelles pour l'imagerie

Outre les présentations, de nombreux exercices encadrés ont été proposés avec la mise en œuvre pratique des algorithmes, dans le langage Python.

Un des enjeux de l'analyse de volumes de données de grandes tailles, multidimensionnelles concerne les méthodes de réduction de la dimension (classiques comme ACP, AFC, MDS, ...) ou issues du « machine learning » (kernel PCA, ...). Cette approche a été abordée lors d'une école thématique qui a eu lieu en 2017. Cette formation, nécessitant des connaissances de base en calcul matriciel, a permis d'approfondir certaines des techniques matricielles (recherche de valeurs propres, décomposition en valeurs singulières), sur le plan à la fois théorique et pratique.

On peut trouver un exemple d'utilisation concrète de ce type de technique présenté lors des JCAD 2018 par Alain Franc, INRA, appliqué à la biologie.

L'exploration de la diversité des protistes : l'apport du calcul intensif

J.-M. Frigerio, P. Chaumeil, F. Rué, S. Thérond, V. Louvet, O. Coulaud & A. Franc [JCAD 2018 - Lyon](#)

De façon un peu générale, toutes ces approches conduisent ou sont la base de certains pans de l'Intelligence Artificielle. De plus en plus d'évènements sont consacrés à ces technologies.

Une introduction sur cette thématique a été réalisée en 2018 dans le cadre des « Journées Système » du réseau ResInfo.

Intelligence artificielle : une longue histoire ... et demain ?

Pierre Gançarski (Université de Strasbourg) [Josy Intelligence Artificielle - Strasbourg 2018](#)

De même, le réseau SARI grenoblois a organisé une journée sur le sujet, avec une présentation de Jean-Luc Parouty particulièrement didactique.

AI Machine Learning & Deep Learning

Jean-Luc Parouty (SIMAP) [Séminaire SARI 2019](#)

Compte tenu de l'engouement engendré autour de l'IA, de nombreuses journées et conférences sont organisées sur le sujet. En particulier, il fait l'objet de sessions spéciales lors des Journées Développement (Jdev) de 2020 et 2017.

5.2.3 Visualisation des données numériques

Un des outils d'analyse les plus utilisés est la visualisation des données. Cependant cette visualisation peut s'avérer particulièrement délicate dans le cadre de très gros volumes de données, et nécessite de s'appuyer sur des solutions techniques spécifiques.

Dans le domaine des données numériques, plusieurs bibliothèques sont particulièrement adaptées aux données de grande taille, ainsi qu'à la visualisation in situ, c'est-à-dire en cours de calcul en ce qui concerne les données de simulation : [VisIt](#) et [ParaView](#). Plusieurs interventions sur ce sujet ont été réalisées dans le cadre d'une [journée dédiée organisée en 2017 par le réseau Calcul](#).

De même, une [action de formation](#) a été complètement consacrée à ce sujet en 2016 par le réseau Calcul. Elle a en particulier abordé les bonnes pratiques concernant la production de données : formats d'archivage, technique d'analyse, cycle de vie ainsi que les outils de visualisation avancés (visualisation in situ, temps réel, web).

La visualisation des données est également au cœur des problématiques des utilisateurs du calcul intensif. Le projet européen PRACE sur le calcul intensif propose des formations spécifiques, en particulier sur les outils de la [visualisation scientifique](#).

5.3 Mettre en place des méthodes d'analyse et des chaînes logicielles

L'analyse des données peut nécessiter la mise en place d'un workflow de traitement utilisant des chaînes logicielles. Il existe des environnements virtuels de recherche (VRE Virtual Research Environment) qui facilitent la mise en place de ces méthodes d'analyse complexes.

[VIP](#), the Virtual Imaging Platform, est un portail qui permet à ses utilisateurs d'accéder simplement à leurs données, de les traiter facilement avec des logiciels préinstallés sur la plateforme. Traitements et données sont distribués sur l'infrastructure EGI (infrastructure de grille de calcul européenne). Pour répondre au besoin d'interopérabilité des données, l'[API CARMIN](#) (API web) est maintenant utilisée par VIP. Cette présentation explique les différentes étapes du fonctionnement du système mis en place.

VIP : towards data interoperability through CARMIN, vidéo

Axel Bonnet, Pascal Wassong, Frederic Cervenansky, Camarasu-Pop Sorina, CREATIS et , Tristan Glatard, Concordia University [JCAD 2019](#), Toulouse.

Virtual Imaging Platform

Sorina Camarasu-Pop, Axel Bonnet, Frédéric Cervenansky, CREATIS, Tristan Glatard, Concordia University [JCAD 2018](#)

[Pangeo](#) est une communauté qui travaille au développement de logiciels et d'infrastructures pour faciliter la mise en œuvre des géosciences, dans le domaine du « Big Data ». Cette communauté développe tout un écosystème d'outils open source pour les géosciences.

Cet écosystème a été présenté lors des [JCAD 2019](#) et [2018](#) :

Analyse de simulations numériques de l'océan en préparation aux missions satellite : cas d'utilisation des outils PANGEO

A. Albert, F. Briol, L. Brodeau, G. Dibarboure, G. Eynard-Bontemps, J. Le Sommer, A. Ponte [JCAD 2019](#), Toulouse.

Jupyter, Dask : traitement distribué simple et interactif en Python sur HPC avec l'écosystème Pangeo

Guillaume Eynard-Bontemps, Centre National d'Etudes Spatiales [JCAD 2018](#), Lyon.

D'autres environnements de management de workflow existent :

WRENCH : Workflow Management System Simulation Workbench

Frederic Suter, Henri Casanova, Rafael Ferreira Da Silva, CC IN2P3 [JCAD 2018](#), Lyon.

enfin, les environnements de notebooks sont des outils de plus en plus utilisés dans le cadre de l'analyse de données. Les notebooks sont des programmes qui contiennent à la fois du texte et du code, dans différents langages (Python, Julia, R, Scala ...), exécutables via une interface web. Ces outils sont de plus en plus couramment utilisés en sciences des données. Jupyter est l'application de notebooks la plus utilisée actuellement.

Plusieurs interventions ont eu lieu sur ce sujet. La première, exposée lors des JCAD 2019, met particulièrement en avant l'intérêt des notebooks pour la reproductibilité.

Towards reproducible Jupyter notebooks

Ludovic Courtès, INRIA [JCAD 2019](#), Toulouse.

La présentation suivante, qui a eu lieu lors des JCAD 2018, expose les services de notebooks proposés par l'infrastructure de grille européenne EGI.

EGI Notebooks : Jupyter as a Service and EGI Check-In AAI

Baptiste Grenier, egi.eu [JCAD 2018](#), Lyon.

Enfin, ce dernier exposé, des JCAD 2018, montre l'utilité des notebooks pour la pédagogie et la formation.

RomeoLAB, le portail web HPC : cas d'utilisation pour la pédagogie et les logiciels à la demande

Arnaud RENARD, Université de Reims Champagne-Ardenne [JCAD 2018](#), Lyon.

5.3.1 Qualité logicielle

Ces chaînes de traitement sont très souvent associées à des instruments complexes et nécessitent ainsi de s'interfacer parfaitement avec le reste de l'instrument. Dans ce contexte, il convient d'assurer le suivi des exigences liées au logiciel, la gestion des interfaces avec le reste de l'instrument et l'activité Assurance Qualité Logiciel. Cette dernière permet notamment de répondre à des exigences applicables à un logiciel, du développement à la maintenance de celui-ci. L'ensemble des activités, normes, contrôles et procédures mis en place doit couvrir la totalité de la durée de vie d'un logiciel. Il est par exemple important de vérifier et valider au travers de tests la bonne santé du code et de constamment veiller à la traçabilité qui lui est liée.

Une [journée du réseau Qualité en Recherche](#) a été entièrement consacrée à ce sujet en 2019. Plusieurs exposés ont permis d'illustrer les concepts associés à la qualité logicielle :

Qu'est-ce qu'un logiciel et qu'est-ce que la qualité ?

Henri VALEINS, Journée thématique Assurance Qualité Logiciel 2019, Paris.

Plans de Gestion de Logiciel et Assurance Qualité Logiciel, les apports de PRESOFT

Geneviève Romier, CC-IN2P3, Journée thématique Assurance Qualité Logiciel 2019, Paris.

Référentiels et normes de codage

Z.Tucsnak, Journée thématique Assurance Qualité Logiciel 2019, Paris.

Qualité Logiciel dans un projet de Nanosatellite

Colin Gonzalez, AstroParticules et Cosmologie, Journée thématique Assurance Qualité Logiciel 2019, Paris.

5.3.2 Retours d'expérience

Dans le domaine environnemental, les chaînes logicielles sont également mises en place pour automatiser et enchaîner un certain nombre de traitements comme :

- le contrôle qualité basé sur des paramètres physiques de l'instrument
- le contrôle qualité spécifique à un type d'instrument
- les corrections
- le filtrage
- les agrégations
- le stockage en base de données

Plusieurs présentations issues des journées du réseau SIST illustrent des mises en œuvre de chaîne logicielle d'analyse de données :

Filtrage interactif de données multidimensionnelles

Patrick Brockmann, Laboratoire des Sciences du Climat et de l'Environnement [SIST16 OSU OREME 2016, Montpellier](#).

Chaînes de traitement en temps quasi réel des mesures de gaz à effet de serre du TGIR ICOS

Lynn Hazan, Laboratoire des Sciences du Climat et de l'Environnement [SIST18 2018, Observatoire Versailles](#).

Vie d'une données sismologique de sa naissance sur le terrain jusqu'à sa distribution

David Wolyniec, OSU Grenoble - Jonathan Schaeffer, OSU Grenoble [SIST16 OSU OREME 2016, Montpellier](#).

5.4 Optimiser l'utilisation des ressources

L'analyse des données peut nécessiter l'usage de ressources assez massives en termes de puissance et de nombre, d'autant plus que la volumétrie des données a tendance à augmenter fortement.

Il est donc important de s'assurer d'optimiser l'utilisation de ces ressources.

En particulier au niveau des codes de calcul, le réseau Calcul propose régulièrement des journées sur ces sujets :

Une journée d'introduction à l'évaluation des performances des codes de calcul

2018, Observatoire de Paris.

Une formation sur l'évaluation de la performance des codes, en particulier à travers les outils Paraver et Scalasca.

2019, Observatoire de Haute Provence.

Enfin, l'optimisation des ressources s'entend aussi au niveau environnemental, afin de réduire l'impact des analyses de données. Le [GDS EcoInfo](#) a organisé en 2017 une [journée sur l'impact des logiciels sur l'environnement](#), dont certains exposés peuvent directement concerner la problématique de l'analyse des données :

Écoconception logicielle, retours d'expérience sur la réduction de l'impact des logiciels

Olivier Philippot, GreenSpector Conf EcoInfo 2017, Grenoble.

TEEC : Logiciel vert et durable

Hayri ACAR, Université Lyon 1 Conf EcoInfo 2017, Grenoble.

Écoconception logicielle pour la gestion des datacentres de calcul

Olivier Richard, LIG Conf EcoInfo 2017, Grenoble.

Calcul Intensif, Consommation et Changement Climatique

Xavier Vigouroux Conf EcoInfo 2017, Grenoble.

Eco-élasticité logicielle pour un Cloud frugal

Thomas Ledoux, Ecole des Mines de Nantes Conf EcoInfo 2017, Grenoble.

5.5 Traitements sémantiques/ linguistiques

L'accroissement massif de la production scientifique (données et publications) et la multitude de canaux de diffusion existants appellent au cœur des activités de recherche, la mise en place de solutions informatiques permettant le repérage, l'extraction, l'exploration et l'analyse de corpus de données. Les méthodes et outils TDM (Text and Data Mining) apportent une aide importante pour explorer et analyser le sens des textes et en donner une représentation compréhensible par les humains et les machines.

En 2013, déjà le réseau Renatis avait accueilli Claire Nedellec et Agnès Girard (INRA) pour illustrer l'usage possible des technologies sémantiques pour la gestion de l'information scientifique et technique ainsi que Fabien Amarger (IRIT IRSTEA) pour témoigner de la construction d'une base de connaissance partant d'un cas d'usage : l'annotation des Bulletins de Santé du Végétal

Dans leur présentation, Claire Nedellec et Agnès Girard expliquent les principes de l'analyse sémantique de texte à travers un exemple en recherche documentaire et présentent le projet TirPhase. La notion d'indexation sémantique à travers un exemple en physiologie animale est abordée en début de présentation sous les traits d'une carte d'identité thématique associée au document. Les auteurs présentent également la notion de termino-ontologie et définissent l'ontologie comme « un graphe où les nœuds sont des concepts et les arcs des relations entre ces concepts ». Elles expliquent que l'analyse sémantique identifie les unités sémantiques du texte et les associe aux concepts de l'ontologie. Partant de là, elle présente le processus de conception de termino-ontologie à partir de corpus en deux étapes : extraction automatique de termes avec l'outil Syntex et structuration et modélisation avec l'outil Protégé. La deuxième partie de la présentation est consacrée à la présentation du projet TriPhase. Ce projet a pour objectif d'analyser les publications d'un département de recherche à des fins stratégiques (analyse quantitative des termes au cours du temps) et disposer d'un moteur de recherche sémantique bibliographique spécialisé. Les auteurs expliquent les différentes phases de la construction de la termino-ontologie TriPhase et l'apport des documentalistes dans ce travail collaboratif.

Fabien Amarger, présente quant à lui un projet qui consiste à construire une base de connaissance à partir de source et de données hétérogènes. Il explique ce qu'est une base de connaissance et comment l'interroger.

Des technologies sémantiques pour l'information scientifique et technique

Claire Nedellec, Agnès Girard, INRAFrédocs2013 - Gestion et valorisation des données de la recherche - 2013, Aussois

Annotation des Bulletins de santé du végétal (BSV) et interrogation

Fabien Amarger , IRIT-IRSTEAFrédocs2013 - Gestion et valorisation des données de la recherche - 2013, Aussois

Plus récemment, Laurence El Khoury (DIST-CNRS) et Stéphane Schneider (INIST – CNRS) à l'occasion des Frédocs2018 ont présenté les projets [ISTEX](#), [VisaTM](#) et [OpenMinteD](#) pour illustrer la mise à disposition d'une infrastructure de text-mining. ISTEEX est une base documentaire qui propose un accès à distance et de manière pérenne à un corpus multidisciplinaire (plus de 23 millions de documents) en texte intégral. Cette base propose également des services permettant la mise en place de traitements des données : extraction, fouille de textes, production de synthèses documentaires. La première partie de l'intervention présente les objectifs, les ressources et les possibilités offertes par la plateforme. La seconde partie s'intéresse plus particulièrement à la fouille de texte et au projet VisaTM. Ce projet porté par l'INRA vise à étudier les conditions de production de services TDM à haute valeur ajoutée basés sur l'analyse sémantique à destination des chercheurs pour une généralisation des approches TDM dans les activités de recherche. Le Projet européen H2020 OpenMinteD présenté en 3e partie est une e-infrastructure encourageant et facilitant l'utilisation des technologies de fouille de textes. Sa connexion à ISTEEX permet l'exploration de corpus. Plateforme open source, ouverte et pérenne elle offre aux chercheurs la possibilité de découvrir, créer, partager et réutiliser des logiciels, des documents et des ressources pour le text-mining, le TALN, l'Extraction d'Information en travaillant à partir de sources documentaires licitement utilisables tels qu'ISTEX, OPENAIRE et CORE.

Bases de ressources numériques et services aux chercheurs. Avec ISTEEX et OpenMintedD, l’alliance pour une infrastructure de text-mining

Laurence El Khoury (DIST-CNRS), Stéphane Schneider (INIST – CNRS)Frédocs2018 - Démarches innovantes en IST : expérimenter, proposer (se) réinventer, 2018, Albi

Préserver et archiver

Préserver, sécuriser l'information et sauvegarder, voire archiver les données sont des phases essentielles de la gestion rigoureuse des données, mais il n'est pas toujours aisé de faire la distinction entre ces notions et d'utiliser le bon terme et la procédure associée. De plus, préserver pour un usage futur dont on ignore le plus souvent les caractéristiques est compliqué. C'est pourquoi des retours d'expériences avec leurs succès et leurs échecs sont intéressants à faire connaître. Ces retours d'expérience sont complétés par des conseils appropriés pour sélectionner les données à préserver et pour mener à bien la préservation d'objets numériques.

6.1 Comprendre et différencier les différents concepts

Les notions de stockage, de sauvegarde et d'archivage ainsi que les actions de préservation et de pérennisation ne sont pas toujours définies dans les mêmes termes. Afin de faciliter la lecture de ce chapitre et aider à distinguer les différences entre les termes utilisés, nous vous proposons les définitions suivantes.

6.1.1 Définitions générales

Stocker : C'est l'étape première qui consiste à déposer les données sur un support numérique pour les rendre accessibles. Cela peut être un ordinateur personnel, un disque partagé ou tout autre organe de dépôt. Le stockage permet d'assurer la continuité de l'exploitation sur du court terme. A ce stade, la donnée n'est ni sauvegardée et ni sécurisée.

Sauvegarder : La sauvegarde consiste à dupliquer les données sur un support numérique externe à celui où elles sont stockées. L'objectif est de pouvoir les retrouver en cas de perte ou de dégradation de l'organe de stockage. Il s'agit d'une sauvegarde octet par octet dans une perspective de court ou de moyen terme. La recherche de la préservation de l'intelligibilité des données n'est pas un élément pris en compte.

Cette étape de sauvegarde doit s'accompagner d'une réelle politique de sauvegarde, qui détermine en fonction de la criticité et de la sensibilité des données combien de copies de sauvegarde on établit par jour, par semaine, par mois. Les sauvegardes se font le plus souvent avec des logiciels spécialisés qui permettent de définir ce qu'on sauvegarde et sa fréquence. Le logiciel permet également de restaurer, c'est-à-dire de rétablir les données d'une certaine sauvegarde choisie. La sauvegarde est mise en place par les administrateurs système et réseaux. Dans le cycle de vie de la donnée, les procédures de sauvegarde doivent être définies lors de la partie *Collecter*

Laurent Pelletier, lors de l'ANF PostgreSQL Administration, a présenté les questions à se poser pour définir sa politique de sauvegarde autour des bases de données.

Sauvegardes

Laurent Pelletier, INIST-CNRSANF « PostgreSQL Administration », réseau rBDD, Sète, 2017

Archiver : L'archivage consiste à ranger un document dans un lieu où il sera conservé pendant une période plus ou moins longue et d'y associer les moyens pour réutiliser les données : la réutilisation se faisant en ajoutant de l'intelligence à la sauvegarde. Le contenu des documents archivés n'est pas modifiable. Par contre le contenant (format) des documents archivés peut être modifié (pour éviter l'obsolescence logicielle).

Le terme archive est défini par le législateur : *les archives sont l'ensemble des documents, y compris les données, quels que soient leur date, leur lieu de conservation, leur forme et leur support produits ou reçus par toute personne physique ou morale, et par tout service ou organisme public ou privé dans l'exercice de leur activité* (art. L. 211-1 du code du patrimoine). Les données de la recherche entrent pleinement dans le périmètre des archives.

Pour en savoir plus sur le statut des archives scientifiques du CNRS et sur leur délai de conservation, nous vous conseillons ces deux documents :

Textes réglementaires et durée de conservation

Marie-Laure Bachèlerie, DAJ-CNRSSéminaire « Archivage Numérique des Données de Recherche », réseau SARI, Grenoble, 2019

Traçabilité des activités de recherche et gestion des connaissances - Guide pratique de mise en place

Alain Rivet, CERMAV & Marie-Laure Bachèlerie, DAJ-CNRS & Auriane Denis-Meyere, IBS & Delphine Tisserand, ISTerreMITI-CNRS, 2018

Préserver : Cette action fait référence au fait de garantir, protéger, mettre à l'abri, sauver d'un dommage ou d'une destruction (cf. notion de sauvegarde) et au fait de tenir dans le même état, en bon état (intelligible). Elle fait aussi référence à la notion de permanence dans le temps (cf. notion d'archivage). Le synonyme « conserver » est utilisé quand il est fait référence à une politique.

Pérenniser : Ce verbe est souvent utilisé à la place de préserver quand on pense archivage pérenne. L'archivage pérenne a pour fonction d'assurer la conservation à long terme des données, leur accessibilité tout en préservant leur intelligibilité, comme rendre accessible en lecture des données immuables (archives de documents administratifs, données de mesures expérimentales, résultats de simulations coûteuses à produire, etc.).

Dans l'article « [l'archivage des données de la recherche à l'Inra. Eléments de réflexion, démarche et perspectives](#) », les auteurs indiquent que pour eux, la pérennisation et la préservation sont le même concept : *La pérennisation (ou préservation) permet de faire face à la perte d'informations d'identification ainsi qu'à l'obsolescence des supports et des logiciels. Elle consiste en effet à identifier et à conserver des documents et des données pour les rendre accessibles sur le moyen (10 ans et plus) et le long terme (50 ans et plus)*

Dans la suite de ce chapitre, nous utiliserons les termes « préserver / préservation » qui sont les termes le plus utilisés actuellement.

6.1.2 Préserver la masse de données scientifiques

Cette présentation est une bonne entrée en la matière. Cristinel Diaconu introduit sa présentation en illustrant le « Big Bang » des données et en questionnant le destin des masses de données scientifiques (*big scientific data*) sur le long terme. Il explique la fragilité des données numériques et la limite dépassée des capacités de stockage des données. Pourtant les données scientifiques sont riches en information, car structurées selon un plan de recherche et une démarche scientifique, elles sont de plus en plus massives et diverses, produites avec des efforts humains et financiers significatifs, elles englobent des connaissances uniques qu'il faut absolument préserver. Partant d'exemples issus de la physique des hautes énergies produits par des instruments gigantesques pour des collaborations internationales, il démontre l'importance de se préoccuper du sort de ces données et de planifier leur préservation. Il aborde la question du risque associé à la réutilisation, de l'organisation sur le long terme, et des différents modèles de données et niveaux de préservation.

La dernière partie de sa présentation est consacrée au projet **PREDON** qui propose une approche nouvelle mélangeant les capacités scientifique, technique et organisationnelle des grandes collaborations en physique des particules et astrophysique pour définir et construire un système robuste de stockage et analyse des données à long terme. Il présente les principaux défis scientifiques associés à ce projet (nécessité d'identifier les demandes et possibilités techniques pour une approche unifiée, besoin de cohérence et standardisation, de mise en place d'un suivi des lots de données, etc), le plan multi annuel, les compétences et challenges dans la préservation des données, la méthodologie de l'archivage au CINES.

La préservation des données scientifiques

Cristinel Diaconu, CPPMFRéDoc « Gestion et valorisation des données de la recherche », réseau Renatis, Aussois, 2013

6.1.3 Protéger et sécuriser le patrimoine scientifique

L'analyse de risque suit un processus normé (ISO 2700x) comme nous le montre Alain Rivet. Cette présentation inclut une étude de cas sur l'utilisation d'un portable par un chercheur qui illustre le passage de la théorie à la pratique.

La sécurité de l'information est définie comme la « protection de la **confidentialité**, de l'**intégrité** et de la **disponibilité** de l'information ». Elle est aujourd'hui une des problématiques majeures de nos unités et s'appuie sur la mise en place des Politiques de Sécurité des Systèmes d'Informations (PSSI) par les services compétents au sein des structures de recherche.

Nous pouvons résumer la finalité de la « protection du patrimoine scientifique » comme étant le fait de :

- garantir la disponibilité de l'outil de travail pour l'ensemble des personnels de la structure ;
- garantir (si besoin) la confidentialité des informations ;
- garantir l'intégrité des informations et des personnes ;
- assurer la protection des données sensibles de la structure (données scientifiques et techniques ; données de gestion administrative, données individuelles) ;
- assurer la protection juridique (risques administratifs, risques pénaux, atteinte à la réputation scientifique ou à l'image institutionnelle).

L'analyse de risques dans le cadre du système d'information

Alain Rivet, CERMAV6ème rencontres du Réseau Qualité en Recherche, réseau QeR, Biarritz, 2015

Cyril Bras ajoute un quatrième critère pour définir le niveau de sécurité d'un élément du système d'information : la **preuve** qu'il définit comme la propriété d'un bien permettant de retrouver, avec une confiance suffisante, les circonstances dans lesquelles ce bien évolue. Cette propriété englobe notamment :

- la traçabilité des actions menées ;

- l'authentification des utilisateurs ;
- l'imputabilité du responsable de l'action effectuée.

Les solutions peuvent être apportées à différents niveaux :

- la sauvegarde avec des outils comme Rsync et Bacula ;
- le chiffrement des données ;
- les droits d'accès ;
- la protection réseau ;
- l'éducation des utilisateurs.

La sécurisation des données (accès, sauvegarde, archivage...)

Cyril Bras, CERMAV 7^{ème} rencontres du Réseau Qualité en Recherche, réseau QeR, Grenoble, 2017

La protection du patrimoine scientifique et technique est l'affaire de tous, mais elle ne se limite pas à des mesures techniques.

Sans prise de conscience de tous sur la nécessité de préserver le patrimoine scientifique, une disparition (partielle sinon généralisée) de nos données est plus que probable. Marion Massol lors de sa présentation au séminaire SIST en 2016 nous rappelle cependant que des solutions d'avenir existent déjà (au CINES).

Patrimoine scientifique en danger : des solutions d'avenir existent déjà

Vidéo : Marion MASSOL, CINESSéminaire SIST 2016 : réseau SIST, Montpellier, 2016

6.2 Préserver les objets numériques

Tous les objets numériques ne nécessitent pas les mêmes opérations pour être préservés. Leur préservation dépend parfois de leur nature (données textuelles, données numériques, données audiovisuelles, modèles et codes informatiques ..), du niveau de leur protection (chiffrés, non chiffrés) de leur façon d'être collectés (observateur, capteur, modèle, etc) ou de leur évolution dans le temps (stable, croissante, révisable).

Voici quelques exemples de préservation d'objets numériques qui ont fait l'objet de présentation.

6.2.1 Les données d'un tableur

Marie-Claude Quidoze revient sur les quatre risques menaçant inéluctablement un fichier sur une longue période et donne des pistes d'améliorations possibles. Elle fait trois recommandations importantes :

- ne pas penser uniquement format ouvert, mais penser aussi format durable pour une sauvegarde à long terme,
- dans le monde des tableurs, où à ce jour les seuls formats durables sont CSV / TXT, ne pas oublier de prendre en compte, lors de la création des fichiers, la perte de fonctionnalités de ce format par rapport aux formats natifs,
- avoir le réflexe de valider son fichier avec l'outil [FACILE](#) du CINES.

Sécuriser les données produites par les carnets de terrain électroniques

Marie-Claude Quidoze, CEFEAtelier « Carnets de terrain électroniques », Montpellier, 2018

6.2.2 Les bases de données

En Avril 2004, le CINES a publié un « [Guide Méthodologique pour l'archivage des bases de données](#) » que nous recommandons fortement, même s'il est un peu ancien (la famille NoSQL est absente). Il contient les bonnes questions à se poser (est-ce une base de données vivante / consultée / cumulative ?), est-ce une base de données pilotée par une interface ? etc). Il présente les différents modes de sauvegarde possibles d'une base de données avec pour chacun leurs avantages et leurs inconvénients. Il liste les différentes documentations à joindre. Et surtout il sensibilise l'utilisateur sur la problématique de l'interface qui du point de vue préservation est un problème à prendre en compte en tant que tel (maillon faible).

En novembre 2014, le réseau rBDD a consacré une journée à cette thématique « [Journée de sensibilisation à la sécurisation et à la pérennisation des données](#) ». À cette occasion, Michel Jacobson a fait une présentation dans laquelle il présente le contexte de la pérennisation des bases de données, le format *Software Independent Archiving of Relational Databases* (SIARD) et un retour d'expérience de l'utilisation de ce format pour la matrice cadastrale numérique.

Retour d'expérience sur l'utilisation du format SIARD pour l'archivage des bases de données relationnelles

Vidéo : Michel Jacobson, LLLJournée « Sensibilisation à la sécurisation et à la pérennisation des données », réseau rBDD, Paris, 2014

6.2.3 Les données chiffrées

Dans cette présentation, François Morris aborde le cas des données protégées par un chiffrement. Après un rappel de ce qu'est le chiffrement, il présente le chiffrement dans la durée : archivage des données chiffrées et utilisation de ces données, donc comment déchiffrer dans le futur ces données archivées.

La pérennisation des données chiffrées ? Quel est l'impact du chiffrement sur le long terme ?

Vidéo : François Moris, CNRSJournée « Sensibilisation à la sécurisation et à la pérennisation des données », réseau rBDD, Paris, 2014

6.2.4 Les données à caractère personnel

Dans cette présentation, Emilie Masson et Patrick Guillot proposent une carte mentale avec deux parties distinctes. À droite du terme « données personnelles et archivage » ils donnent une définition des données à caractère personnel et des traitements de données personnelles. À gauche du terme « données personnelles et archivage », ils rappellent les obligations de la RGPD en termes de conservation après l'exercice de la finalité et ils établissent des recommandations en fonction de trois types d'archives : archives courantes / archives intermédiaires et archives définitives.

Données personnelles et archivage

Emilie Masson, SPD-CNRS & Patrick Guillot, Université Grenoble AlpesSéminaire « Archivage Numérique des Données de Recherche », réseau SARI, Grenoble, 2019

6.2.5 Les logiciels / les codes sources

La préservation des logiciels et des codes sources est indispensable pour assurer la reproductibilité de la science. Le projet Software Heritage a pour objectif de collecter, préserver et rendre disponible le code source (et son historique) de tous les logiciels publiquement disponibles. Une présentation du projet est disponible dans la section *Infrastructures* de ce guide.

6.3 Archiver les objets numériques

L'archivage des objets numériques s'appuie sur les mêmes concepts que la préservation, mais elle demande de choisir un centre d'archivage homologué aux fonctionnalités d'interrogation moindres.

6.3.1 Archivage à long terme et politiques de préservation

Ces retours d'expériences recueillis lors du séminaire « *Archivage Numérique des Données de Recherche* » (Grenoble, novembre 2019) illustrent les démarches en cours dans certains instituts et sur certaines infrastructures pour mettre en oeuvre des politiques d'archivage et de préservation des données.

Ces présentations offrent un panorama des différents processus ou démarches d'archivage à long terme mis en place dans diverses disciplines scientifiques. On y découvre les politiques de gestion ou de partage des données et les démarches spécifiques à chaque structure pour mettre en place ou organiser un système d'archivage et de préservation des données sur le long terme.

Pour commencer, Michel Jacobson présente l'offre de service de la Très Grande Infrastructure de Recherche (TGIR) Huma-Num pour l'archivage à long terme. Ce service co-construit en partenariat avec le CINES consiste en un accompagnement méthodologique et parfois technique pour l'archivage au sein même du CINES. Il prend la forme aussi de recommandations, d'un dialogue, d'une médiation entre les producteurs de données et le CINES et assure la prise en charge des coûts d'archivage pour l'utilisateur. Il illustre sa présentation par différents exemples d'archivage en SHS.

Les présentations de Yonny Cardenas et de Jean-François Perrin présentent les politiques de gestion, partage et diffusion des données au centre Paul Langevin et au centre de calcul de l'IN2P3. Yonny Cardenas insiste en particulier sur les qualités et faiblesses de la politique de gestion mise en place. Il présente un cas d'utilisation qui mobilise des compétences multidisciplinaires de chercheurs, informaticiens et documentalistes et révèle un besoin crucial d'archivage à long terme.

Gilles Duvert témoigne quant à lui de la préservation des données en Astronomie, discipline dans laquelle les données et l'interopérabilité sont au cœur de la démarche scientifique depuis 40 ans. Il signale entre autres la pratique qui consiste à archiver et documenter les données à l'aide de standards partagés au niveau international.]

Archivage des données à Huma-NUM

Vidéo : Michel Jacobson, Huma-NumArchivage Numérique des Données de Recherche, 2019, Grenoble

Archivage des données à l'IN2P3

Vidéo : Yonny Cardenas, CC-IN2P3Archivage Numérique des Données de Recherche, 2019, Grenoble

Gestion des données scientifiques à l'Institut Laue Langevin

Jean-Francois Perrin, ILLArchivage Numérique des Données de Recherche, 2019, Grenoble

L'archivage des données en astronomie

Vidéo : Françoise Genova, observatoire astronomique de Strasbourg, Gilles Duvert, IPAG, OSUG et JMMC Archivage Numérique des Données de Recherche, 2019, Grenoble

6.3.2 Les plateformes d'archivage des données

Les plateformes d'archivage sont décrites dans la partie *Infrastructures*.

6.4 Sélectionner les données pertinentes

Cette étape nécessite une phase de sélection des informations pertinentes (validées, utiles...) pour son activité tout en se préoccupant de leur exploitation future à travers les problématiques de durée de vie, de confidentialité et de sécurité des données.

Plusieurs critères sont à prendre en considération :

- la date et la fréquence à laquelle faire cette sélection : fin de thèse, de projet ou de contrat / avant de quitter son emploi / à date régulière / etc ;
- la durée de conservation : durée officielle pour les documents administratifs, à définir en fonction des besoins pour les données scientifiques ;
- l'obligation administrative de destruction éventuelle ;
- la nécessité d'anonymisation éventuelle ;
- le format, la nature des données qui définiront leur lisibilité dans le temps ;
- les supports d'enregistrement, d'utilisation et de stockage des données ;
- la criticité et donc le niveau de sécurité et d'accessibilité nécessaires pour protéger les données ;
- le coût de ces supports ou encore des modifications de format de fichiers, ou bien de l'espace de stockage nécessaires à la conservation des données.

Ses critères ont été établis dans le cadre de l'archivage pérenne, mais ils s'appliquent tout à fait à l'archivage (nota bene : la différence entre archive pérenne et archive porte sur la durée de conservation).

Dans la présentation référencée, Lorène Béchard et Marion Massol présentent le Centre Informatique National de l'Enseignement Supérieur (CINES), ses missions et l'infrastructure d'archivage du CINES. Elles abordent les différents formats, le cas particulier de l'archivage des bases de données et les questions à se poser avant d'archiver. Elles précisent que l'archivage pérenne des documents électroniques consiste à conserver le document et l'information qu'il contient :

- dans son aspect physique comme dans son aspect intellectuel,
- sur le très long terme et au-delà,
- de manière à ce qu'il soit en permanence accessible et compréhensible.

Les critères à prendre en compte pour la conservation des données

Vidéo : Lorène Béchard, CINES & Marion Massol, CINES Journée « Sensibilisation à la sécurisation et à la pérennisation des données », réseau rBDD, Paris, 2014

Dans cette présentation, après avoir évoqué le contexte réglementaire en matière de gestion d'**archives publiques**, Magalie Moysan, coordinatrice du pôle Sécurisation des données et documents et responsable du département archives à l'Université de Paris, aborde les enjeux de leur préservation ainsi que les outils et méthodes disponibles pour procéder à leur sélection. Elle insiste sur le statut « d'archives publiques » souvent méconnu dans nos établissements et sur les obligations légales et réglementaires qui y sont associées notamment l'obligation de bien gérer, conserver et archiver ses documents ou l'interdiction de les détruire sans visa préalable des autorités compétentes. Elle explique ensuite

l'intérêt de concevoir l'archivage comme un moyen de « conserver les preuves », de garantir une antériorité, d'assurer le suivi d'une activité, de fiabiliser les résultats et de constituer un patrimoine scientifique. La sélection des données en vue de leur conservation suppose très en amont de pouvoir anticiper cette tâche en procédant à une description rigoureuse des données et de leur contexte de production (contenu, date, modalité d'entrée, conditions d'accès etc.). Le processus de sélection s'organise autour de quelques grands principes (intérêt scientifique, juridique, historique de la donnée, intelligibilité de la donnée) et se détermine souvent dans le cadre d'un échange entre scientifiques et archivistes. Les référentiels de conservation sont des outils d'aide intéressants pour accompagner le processus de sélection.

Sélectionner les données pour la préservation : enjeux et méthodes

Magalie Moysan, Université de Paris Atelier Dialogu'IST « Rendre FAIR les données, mais quelles données préserver ? », réseau Renatis, 2020

6.5 S'appuyer sur les enseignements des retours d'expérience

Le premier retour d'expérience du siècle dernier est un échec tandis que le deuxième datant du début du XXI^e siècle est synonyme de succès.

6.5.1 Retour d'expérience sur des données numériques en Écologie

À travers ce retour d'expérience, Marie-Claude Quidoz explique pourquoi, des données pourtant préservées avec attention au CEFÉ en 1984, ont été impossibles à réutiliser vingt-cinq ans plus tard. Cette expérience malheureuse est riche d'enseignements et permet très rapidement de comprendre sur un exemple concret les difficultés rencontrées qui sont de quatre ordres :

- l'obsolescence matérielle;
- l'obsolescence logicielle;
- l'obsolescence du format du fichier;
- la perte de signification du contenu.

Retour d'expérience - Les données de l'Écothèque Méditerranéenne

Vidéo : Marie-Claude Quidoz, CEFÉ Journée « Sensibilisation à la sécurisation et à la pérennisation des données », réseau rBDD, Paris, 2014

6.5.2 Retour d'expérience sur des données orales en SHS

Michel Jacobson débute sa présentation par la présentation de la plateforme technique Cocoon pour « Collections de CORpus Oraux Numériques ». C'est une plateforme technique qui accompagne les producteurs de ressources orales, à créer, structurer et archiver leurs corpus ; un corpus pouvant se composer d'enregistrements (en général audio) accompagnés éventuellement d'annotations de ces enregistrements. Cette plateforme s'appuie sur les services d'Huma-Num. Ensuite, il détaille les différentes étapes suivies pour mener à bien l'archivage pérenne mis en place en collaboration avec le CINES.

Organisation du Centre de ressources numériques Cocoon comme service versant pour l'archivage de données orales en SHS

[vidéo](#) Michel Jacobson, Journée « Sensibilisation à la sécurisation et à la pérennisation des données », réseau rBDD, Paris, 2014

Cette dernière étape d'un projet de recherche représente en quelque sorte la finalité de toute une politique de gestion de données puisqu'elle vise, dans un contexte de Science ouverte, à publier et diffuser les données de manière à ce qu'elles soient bien accessibles et réutilisables selon des formats et des processus interopérables.

L'accompagnement des réseaux métiers s'exerce sur le processus de publication des données dans des entrepôts ou des plateformes techniques, pour en permettre l'accès, ainsi que sur la documentation des données avec des métadonnées descriptives provenant de vocabulaires contrôlés et de leurs formats d'exploitation pour en assurer la réutilisabilité. Ainsi, les réseaux travaillent sur l'ensemble des informations (données, métadonnées, modes opératoires, échantillons, publications, visualisation et interfaces graphiques) nécessaires à la mise en œuvre des supports de diffusion et de valorisation les plus pertinents en rapport avec l'objectif du projet initial.

Cette étape de publication et de diffusion est en outre accompagnée désormais d'une action nécessaire d'identification des données via des identifiants pérennes lors du dépôt dans des entrepôts de données.

7.1 Communiquer et documenter

Finaliser le Plan de Gestion de Données

Pour rappel, en fin de projet, il ne faut pas oublier de mettre à jour et finaliser la rédaction du plan de gestion de données. Il est nécessaire de s'assurer que les premières informations saisies sont encore valides et effectuer une mise à jour en ajoutant les dernières informations disponibles.

7.2 Publier les métadonnées

Utilisation de catalogues de métadonnées

Les catalogues de métadonnées représentent un moyen cohérent et rigoureux pour décrire et publier des jeux de données. Ils permettent de faciliter la recherche et l'identification des données (F de FAIR).

Pour être interopérables, ces catalogues s'appuient en général sur des normes pour représenter les métadonnées. Par exemple, dans les sciences de l'environnement les normes ISO 19115 et ISO 19139 sont des références pour représenter l'information géographique à l'aide de métadonnées dans les domaines où les données sont géospatialisées.

- L'ISO 19115-1 définit le schéma requis pour décrire des informations géographiques et des services au moyen de métadonnées. Elle fournit des informations concernant l'identification, l'étendue, la qualité, les aspects spatiaux et temporels, le contenu, la référence spatiale, la représentation des données, la distribution et d'autres propriétés des données géographiques numériques et des services.
- L'ISO 19139 définit le schéma d'implémentation et d'encodage XML pour représenter les métadonnées ISO 19115.

En ce sens, dans le domaine environnemental où les données proviennent fréquemment de mesures géolocalisées sur le terrain, le logiciel [GeoNetwork](#) est une des références importante pour décrire et représenter les jeux de données géolocalisés et constituer un catalogue qui inventorie les différents jeux de données d'un Institut.

Grace a l'utilisation de protocoles d'échanges normalisés, ce logiciel permet d'interagir avec d'autres catalogues de ressources spatialisées via le [protocole CSW de l'OGC](#) (Catalogue Services for the Web) et permet ainsi de construire un réseau de catalogues interagissant les uns avec les autres, réseau notamment demandé par la [Directive Européenne Inspire](#).

Le logiciel GeoNetwork utilise, en outre, le protocole [WMS \(Web Map Service\) de l'OGC](#), pour pouvoir interagir avec des serveurs cartographiques comme [GeoServer](#), ce qui permet de représenter les points de mesures sur une carte et de pouvoir accéder aux données brutes associées à un point de mesure géoréférencé.

L'utilisation des logiciels GeoNetwork et GeoServer a été traitée par une action de formation ANF du réseau SIST pour laquelle on trouvera les informations sur les [pages formation du site SIST](#)

Documentations sur les logiciels étudiés GeoNetwork et GeoServer

J. Fabre, M. Libes, O. Lobry, D. Mallarino, M. Rouan, J. Schaeffer ANF SIST 2017 Fréjus

Le recueil des métadonnées ainsi que la rédaction et la mise à jour des métadonnées dans des fiches adaptées sont souvent ressentis comme contraignants. Cependant le logiciel GeoNetwork propose une interface de programmation (API) qui permet d'automatiser la constitution des catalogues. Plusieurs développements se sont intéressés à l'utilisation de l'interface de programmation (API) de Geonetwork pour pouvoir insérer automatiquement des métadonnées dans les fiches avec des programmes écrits en langage « R ».

C. Bernard J. Fabre, et O. Lobry indiquent comment alimenter un catalogue de données GeoNetwork de l'OSU Oreme, de manière automatique à partir de données stockées dans une base de données interne à l'unité.

Mise en place de catalogues INSPIRE et de leur alimentation automatique

C. Bernard, J. Fabre, et O. Lobry [Séminaire SIST16 Montpellier](#)

De la même manière, Emmanuel Blondel est l'auteur d'un ensemble de bibliothèques de programmation écrites en « R », destinées à faciliter l'insertion de métadonnées dans les catalogues « GeoNetwork ». Ces développements ont été présentés lors d'un atelier organisé par le réseau RBDD et SIST :

- [Atelier « Métadonnées et R »](#)
 - Écrire et Lire des métadonnées avec la librairie R *geometa*
 - Gérer des données dans GeoServer avec la librairie R *geosapi*

— Gérer des métadonnées dans GeoNetwork avec la librairie R *geonapi*

GeoFlow : workflow R pour gérer les données spatiales

Julien Barde, Emmanuel Blondel et Wilfried Heintz [Séminaire SIST19 Toulouse](#)

Ce logiciel de catalogage est utilisé dans de nombreux portails de données comme le catalogue Sextant d'Ifremer pour lequel M. Treguer nous indique les services de l'OGC utilisés.

Visualisation et analyse des données marines et littorales avec Sextant : Utilisation des services OGC

Michael Treguer [Séminaire SIST16 Montpellier](#)

A titre d'exemple, de nombreux OSU (Observatoire des Sciences de l'Univers) mettent en oeuvre ces catalogues « geonetwork » très utiles pour inventorier les jeux de données disponibles dans les unités de recherche :

- [portail de données OSU Oreme](#)
- [portail de données Indigeo](#)
- [portail de données OSU Pytheas](#)

B. Debray nous présente le projet DataOSU à l'Observatoire de Besançon destiné à élaborer un [portail de données original propre à l'OSU Theta](#). Il décrit toute l'organisation et le développement nécessaire à la réalisation du projet. La nécessaire collecte des métadonnées auprès des chercheurs et le mapping sémantique destiné à assurer l'interopérabilité avec les standards existants du Dublin core, IVOA, Datacite, GBIF

Le projet Dat@OSU de gestion et valorisation des données de la recherche

Bernard Debray, Univers, Transport, Interfaces, Nanostructures, Atmosphère et environnement, Molécules [Séminaire SIST16 Montpellier](#)

7.2.1 La directive européenne « INSPIRE »

La directive **INSPIRE**, élaborée par la Direction générale de l'environnement de la Commission européenne en 2007, vise à établir en Europe une infrastructure de données géographiques pour favoriser la protection de l'environnement, assurer l'interopérabilité entre bases de données et faciliter la diffusion, la disponibilité, l'utilisation et la réutilisation de l'information géographique en Europe. INSPIRE vise ainsi à mieux partager les données de la recherche.

Marc Leobet, chargé de mission à la Mission information géographique du ministère en charge du développement durable pose, dans cette présentation réalisée en 2013, le cadre de la Directive Inspire. Il présente tout d'abord l'utilité de cette Directive (identification des données, gestion de la confidentialité, les problèmes de conventionnement et la qualité des données), son contexte, les obligations qu'elle induit, le contexte autour de la réutilisation des données du secteur public et l'application de la Directive inspire dans le domaine de la recherche.

La Directive INSPIRE pour les néophytes

F. Merrien, M. Léobet, M. Francès Mission de l'information géographique du ministère de l'Environnement

Gestion et valorisation des données de la recherche

Marc Leobet, Chargé de mission et PCE INSPIRE [Frédocs2013 -7](#) au 10 octobre 2013, Aussois

7.2.2 Les portails nationaux et européens

Les pôles de données nationaux en environnement

Les catalogues et portails d'accès aux données se matérialisent également au niveau national où une infrastructure de recherche (IR) Data terra a été créée pour accéder aux données spatiales et in situ du système Terre.

La mission de l'IR Data Terra consiste à organiser de manière intégrée la diffusion et l'accès aux données, en mettant à disposition les données, les produits et des services relatifs à l'observation du système Terre, via les pôles de données et de services existants :

Présentation de l'IR Data terra

Richard Moreno, directeur technique IR Data Terra Séminaire SIST 2019, OMP Toulouse

l'IR Data Terra est constitué de quatre pôles de données (ODATIS, AERIS, ForM@Ter & Theia dont la mission principale est de mettre à disposition des données, des produits, des logiciels, des outils et/ou des services destinés en premier lieu à la communauté scientifique française dans le cadre de ses recherches sur le système Terre.

Les informations proposées par les pôles de données sont aussi fondamentales pour la mise en œuvre des politiques publiques. En permettant de mieux comprendre la structure et le fonctionnement du système Terre, les travaux utilisant ces données ont un impact socio-économique important dans des domaines tels que les risques naturels, le changement climatique, les ressources minérales ou les ressources en eau. Dans ce contexte, les pôles servent aussi la communauté internationale (missions satellites, réseaux d'observation internationaux, partenariats pour le développement).

Les politiques opérationnelles de ces pôles sont suivies par le réseau SIST où elles ont été présentées :

- Données océanographiques : [Pôle Odatis, supports de présentation](#).
- Données atmosphériques : [Pôle Aeris, supports de présentation](#).
- Données terre solide : [Pôle Form@ter, supports de présentation](#).
- Données surfaces continentales : [Pôle Theia, supports de présentation](#).

Les portails de données européens

Par ailleurs, de grands projets européens mettent désormais en place des portails d'accès à très large échelle. C'est le cas du portail du [projet Seadatanet](#) qui vise à rassembler les données marines de plus de 30 pays européens. L'intérêt de ces portails est de fournir toutes les garanties d'une interopérabilité maximale basée sur des protocoles standards et des thesaurus et vocabulaires contrôlés du [British Oceanographic Data Center : BODC](#). Le projet européen Seadatanet vise à élaborer et mettre en place un portail européen d'accès aux données marines en se basant sur de nombreux standards rendant les données FAIR.

Seadatanet est un exemple d'envergure européenne pour la mise en place de standards d'interopérabilité. Il repose sur de nombreux vocabulaires contrôlés fournis par le BODC. Une présentation du projet Seadatanet a été faite par Michele Fichaut et Florence Conquet

Présentation du projet SeaDataNet, interopérabilité à l'échelle paneuropéenne

Michèle Fichaut, Systèmes d'Informations Scientifiques pour la Mer Séminaire SIST15 OSU Pytheas Marseille 2015

Soumaya Lahbib au séminaire SIST18 à l'observatoire OVSQ de Versailles, présente un exemple de dépôt de données de cytométrie en flux dans le portail de données Seadatanet. Il est intéressant de prendre connaissance de la démarche et du workflow de traitement nécessaire pour intégrer des données dans un portail interopérable qui respecte un grand nombre de standards.

Interopérabilité des données issues d'analyses par Cytométrie en Flux dans l'infrastructure européenne Sea-DataNet

7.3 Utilisation de thesaurus

Un vocabulaire contrôlé est une liste de termes (mots et expressions) soigneusement choisis pour désigner les concepts d'un domaine (un seul terme préférentiel et éventuellement plusieurs entrées non préférentielles). Ces vocabulaires sont regroupés dans des « thesaurus » qui sont des listes organisées de termes, contrôlés et normalisés, (descripteurs et non-descripteurs) représentant les concepts d'un domaine de la connaissance.

Un thesaurus permet donc d'organiser et de structurer un vocabulaire d'un domaine de connaissances à partir de relations sémantiques entre concepts (de types hiérarchiques ou associatifs) et d'équivalence entre termes. Il réduit donc l'ambiguïté inhérente au langage humain naturel dans lequel différents noms peuvent être attribués à un même concept.

De nombreux thesaurus existent dans divers domaines scientifiques. Par exemple, dans le domaine environnemental, on utilise fréquemment les thesaurus :

- « *Inspire* » ou
- « *GEMET* ». Ce dernier est un thesaurus documentaire multilingue développé et publié par l'Agence européenne pour l'environnement.

Cependant selon le domaine scientifique et dans certaines disciplines, lorsque les standards, thesaurus et vocabulaires contrôlés n'existent pas, ils doivent alors être créés. Les communautés scientifiques peuvent alors se saisir d'outils tels que "*opentheso*" et thesauform pour répondre aux besoins de normalisation.

Ainsi, lors du séminaire SIST 2018, Dominique Vachez a présenté, en s'appuyant sur le thesaurus *T-Semandiv*, les conditions requises pour une interopérabilité sémantique dans le domaine de la biodiversité : choix de vocabulaires contrôlés et structurés en relations sémantiques utilisés comme référentiels permettant le partage et le croisement des données/métadonnées.

T_Semandiv le thesaurus de la biodiversité

Poster résumant le thesaurus *T_Semandiv* Dominique Vachez, Institut de l'information scientifique et technique Séminaire SIST18 à l'Observatoire de Versailles, 2018

La première version de ce thesaurus a été élaborée avec l'outil *ThesauForm* développé par Baptiste Laporte. *ThesauForm* est un outil pour faciliter la création d'un thesaurus collaboratif. Ces deux points forts sont une élaboration collaborative des termes et une procédure de vote. Cet outil a été utilisé pour construire le thesaurus *T-SITA* qui est le fruit du groupe de travail « CESAB/BETSI ». Ce thesaurus a été utilisé pour annoter des données dans leur base de données à partir du vocabulaire créé.

Création d'un thesaurus collaboratif : cas d'un groupe CESAB, Fondation pour la Recherche sur la Biodiversité, 2015

Baptiste Laporte (Centre de synthèse et d'analyse sur la biodiversité) JrBDD 2015, Sète, mercredi 21/10/2015

Thesauform un outil collaboratif pour faciliter la création de vocabulaire contrôlé par des experts de domaine

MC Quidoz Séminaire SIST18 à l'OVSQ de Versailles

En Archéologie, Blandine Nouvel nous présente l'intérêt du thesaurus *PACTOLS* pour l'archéologie sur le web des données de manière à ouvrir son utilisation au-delà des seules bibliothèques.

Évolution et nouvelles pratiques autour du thésaurus PACTOLS de Frantiq pour l'édition numérique en archéologie

Blandine Nouvel (Centre Camille Jullian / Frantiq)

La révision des PACTOLS au regard du Backbone Thesaurus Blandine Nouvel (Centre Camille Jullian / Frantiq)

Enfin, notons la création du logiciel open source [opentheso](#) qui permet l'élaboration collaborative d'un thésaurus tout comme ThesauForm, mais aussi la gestion de thésaurus multilingue supportant la polyhiérarchie, en conformité avec la norme ISO 25964.

7.4 Utilisation d'identifiants pérennes

Afin d'être cités et réutilisés, les données et documents numériques se doivent de disposer d'un identifiant pérenne pour qu'ils puissent être référencés, visibles et accessibles de manière univoque.

Il existe différents types d'identifiants pérennes pour toutes sortes d'objets y compris les humains. Cet article de J-L Archimbaud fait le point sur les identifiants des documents numériques et leurs usages :

Identifiants des documents numériques : ISBN, ISSN, URL, Handle, DOI, OpenURL, OAI, ARK

Jean-Luc Archimbaud Journées « Conduire et construire un plan de gestion des données : de la base de données à la pérennisation » du réseau CNRS Bases de Données (rBDD) Sète – 22 oct 2015

Il faut aussi noter que dans le domaine de la bio-informatique, des identifiants uniques sont attribués aux enregistrements de séquences DNA ou de protéines. Ils sont nommés [accession number](#).

7.4.1 Les DOI : « Digital Object Identification »

Dans le domaine des données, les D.O.I (Digital Object Identification) DOI est un identifiant pérenne favorisant le référencement et la citation des jeux de données. Ils permettent de citer un jeu de données homogène de manière univoque et de les lier aux publications ou à tout autre produit de recherche. Ils concourent donc à l'identification, la traçabilité et à l'interopérabilité des données. Ils garantissent un lien stable à la ressource en ligne et font correspondre en permanence l'identité de la ressource à sa localisation sur le web.

L'INIST du CNRS est l'agence d'attribution de l'identifiant DOI en France pour l'Enseignement Supérieur et Recherche (ESR). Les D.O.I sont obtenus auprès de l'organisme international « [Datacite](#) ».

L'allocation de D.O.I sur des données implique des devoirs de la part du déposant, qui est de maintenir un lien permanent vers les données identifiées pendant une certaine durée, à travers une page de description (appelée aussi « *landing page* ») qui permet de fournir les métadonnées principales pour décrire les données et y accéder.

Pour créer une « landing page », page d'accueil pour décrire un jeu de données, il faut s'assurer que certaines métadonnées obligatoires sont bien mentionnées et renseignées pour permettre une recherche. Le site [Datacite](#) rappelle quelles sont les [métadonnées obligatoires](#). Pour en savoir plus sur les identifiants pérennes, on peut consulter la [page de Doranum](#)

Attention la pérennité demandée est purement une question de service et n'est pas inhérente à un objet, ni conféré par une syntaxe de nommage particulier. Maintenir la pérennité du lien vers la localisation de la ressource est de la responsabilité du déposant ou du créateur de l'identifiant.

Pourquoi citer les données ?

Herbert Gruttemeier illustre ses propos par des exemples de jeux de données exposés et cités dans différents entrepôts. Il présente la position « officielle » des éditeurs sur l'accès aux données de la recherche et s'attarde sur le type de données et de ressources concernées par l'attribution de DOI.

Data Cite propose un certain nombre de services (création de différents formats de citation pour les DOI, exposition des métadonnées, schéma de métadonnées DataCite et un environnement de test) que l'auteur détaille. Il est question aussi de « Data Citation Index » et de métrique, de l'importance d'accéder à la découverte des données (principe de moissonnage des métadonnées DataCite), des partenariats avec ORCID, OPENAIR, CODATA, FORCE 11, RDA. . .

DataCite : identifiants pérennes pour le partage des données

Herbert Gruttemeier, INIST/ CNRSFrédocs2013 - Gestion et valorisation des données de la recherche - 7 au 10 octobre 2013, Aussois

Cette présentation est consacrée au service proposé par DataCite. Herbert Gruttemeier explique pour commencer ce qu'est un DOI, le principe de citation, pourquoi utiliser un DOI, comment le DOI s'inscrit dans le système Handle. Il aborde la question de la qualité des DOI qui nécessite la mise en place d'une politique institutionnelle. La suite de son exposé est consacrée à la présentation de DataCite, Consortium international porté par des institutions locales, créé officiellement à Londres en décembre 2009. Il présente les 26 membres, la structure, les différents rôles qui lui sont assignés (agence d'attribution de DOI et agence de donnée).

Pour en savoir plus sur le DOI de DataCite :

DOI de DataCite : Système d'identification pour valoriser les données de la recherche,

Mohamed Salah Yahia INIST

Nécessité de publier en identifiant les jeux de données par des « DOI » : [présentation vidéo sur les DOI de Datacite](#)
Mohamed Salah Yahia, Institut de l'information scientifique et technique du CNRS [Séminaire SIST16 OSU Oreme Montpellier](#)

7.4.2 Comment obtenir des DOI ?

Une unité CNRS a la possibilité de souscrire un contrat avec l'INIST du CNRS pour être détenteur d'un préfixe de DOI qui servira à construire et déposer un DOI.

Workflow d'attribution de DOI par l'Inist-CNRS

Mohamed Salah Yahia [Séminaire SIST16 OSU Oreme Montpellier](#)

Cependant dans le paysage national actuel des données environnementales, certaines infrastructures de recherche comme Data Terra seront en charge de fournir des DOI selon les disciplines concernées. Dans le domaine marin le [pôle de données Odatis](#) fournit d'ores et déjà un service de fourniture de D.O.I via le site [Seanoe](#).

Il est pratique d'avoir un outil logiciel qui vérifie de façon autonome les métadonnées requises pour obtenir un DOI et fasse la demande directement auprès de Datacite. C'est le cas du logiciel Geonetwork, que nous avons vu précédemment, pour élaborer des catalogues de jeux de données. Dans cette présentation Annick Battais indique comment assigner un D.O.I a des jeux de données en utilisant le logiciel de catalogage Geonetwork.

Utilisation d'un outil de catalogage normalisé ISO19139 comme GeoNetwork pour constituer une « landing page » pour un D.O.I

Annick Battais Séminaire SIST2019 Toulouse 2019

7.4.3 Retour d'expériences d'utilisation de DOI

Philippe Techné nous indique comment il fournit des DOI sur des données océanographiques grâce à un contrat passé avec l'INIST du CNRS qui, en tant que membre de DataCite, peut fournir et attribuer des DOI. Il passe en revue les métadonnées obligatoires et la landing page qui est constituée.

Mise en place d'un DOI sur les données d'un réseau d'observations océanographiques

Philippe Techné, Laboratoire d'études en Géophysique et océanographie spatiales [Journée SIST16 Montpellier](#)

Création de DOI sur les données et produits grillés du Service National d'Observation SSS

Philippe Techné, Laboratoire d'études en Géophysique et océanographie spatiales [Journée SIST18 OVSQ](#)

Juliette Fabre et Olivier Lobry nous indiquent leur solution pour attribuer des DOI aux jeux de données du Service National d'Observation « Karst ».

Retour d'expérience sur l'attribution de DOI à l'OSU OREME.

Juliette Fabre, OSU OREME - Olivier Lobry, OSU OREME [Journée SIST16 Montpellier](#)

- Établissement de DOI sur des requêtes dynamiques sur des Bases de données Dans l'atelier traçabilité organisé par RBDD en novembre 2018, MC Quido avait traité la possibilité de mettre un identifiant pérenne sur une requête SQL vers une base de données, pour la rejouer. C'est d'ailleurs une des [recommandations de RDA](#).

identifiant pérenne sur une requête SQL vers une base de données

MC Quido, atelier traçabilité RBDD 2018

Sophie Pamerlon rappelle les définitions des identifiants uniques et persistants, puis présente le « Integrated Publishing Toolkit » (IPT) mis en place par le GBIF Global Biodiversity Information Facility) dans le domaine de la biodiversité et ses nouvelles fonctionnalités, en particulier l'attribution de DOI lors de la publication d'un jeu de données.

Le GBIF et les identifiants persistants : Application des DOI aux jeux de données

Sophie Pamerlon (Système mondial d'information sur la biodiversité - Global Biodiversity Information Facility), 2015RBDD

7.5 Les entrepôts de données

Dans un contexte de science ouverte, les acteurs de la recherche s'accordent aujourd'hui pour considérer les données de la recherche comme des produits de la recherche et appellent à mieux les gérer et à les partager. Le partage des données et des connaissances, mais également le partage des logiciels, des méthodes et des processus n'ont de réel bénéfice que s'ils sont accompagnés en amont par une gestion rigoureuse et de qualité des données, basé sur des principes clairs et consensuels.

Les enjeux liés à la gestion et au partage des données de la recherche nécessitent des outils appropriés communément appelés « Entrepôts de données ». Mais qu'est-ce qu'un entrepôt de données et quelles en sont les principales caractéristiques ? Comment les entrepôts de données contribuent-ils à la gestion et au partage des données ?

Qu'est-ce qui différencie un entrepôt de données d'une base de données classique dans le contexte de l'ouverture des données ? Quels services peut-on attendre d'un entrepôt de données aux différentes étapes du cycle de vie de la donnée ? Comment trouver et choisir un entrepôt de données ? Un certain nombre de ces questions relatives aux entrepôts ont été abordées lors d'une [journée de type Hackaton intitulée « entrepôts de données, comment améliorer le dépôt et le partage des données de la recherche ? »](#). Cette journée consacrée aux entrepôts de données a permis de cerner les fonctionnalités que l'on se doit d'attendre d'un entrepôt de données FAIR et les conditions d'utilisation de ce type de service.

Au terme de cette journée, un document « FAQ » a été rédigé répondant aux questions les plus fréquentes que l'on se pose sur le dépôt de données. Cette FAQ est la [synthèse des échanges](#) qui se sont tenus lors de l'hackathon « Comment améliorer le dépôt et le partage de données de recherche ? ». Elle est enrichie régulièrement des discussions sur la liste "données".

On y répond à des questions fréquentes concernant les entrepôts comme :

- Qu'est-ce qu'un entrepôt de données ?
- Comment choisir un entrepôt ?
- Quels sont les critères à prendre en compte pour sélectionner un entrepôt ?
- etc.

Laurent Pelletier de l'INIST, dans une présentation générale sur les entrepôts de données, revient sur les différentes définitions des données, les métadonnées et les principes FAIR. Il explique pourquoi et comment partager les données et comment les entrepôts de données sont impliqués dans ce partage. Il présente les différents types d'entrepôts, les différentes fonctionnalités et les critères de choix pour un entrepôt.

Les entrepôts de données,

Laurent PELLETIER, INIST ANF rBDD du 5 au 7 novembre 2018 à Sète

Dans cette présentation complète, Jean-Christophe Desconnets passe en revue les rôles, les fonctionnalités et les domaines d'utilisation des entrepôts de données :

Les entrepôts de données : Ou comment rendre les données trouvables, accessibles et réutilisables ?

Jean-Christophe Desconnets [Séminaire SIST2020 OSU Lyon](#)

Les entrepôts de données : pierre angulaire du partage des données de la recherche

Ester Dzale Yeumo (INRA) [participer à l'organisation du management des données de la recherche](#) : gestion de contenu et documentation des données - 2016 Paris

Les entrepôts de données de recherche

Sylvie Cocaud (INRA) Participer à l'organisation du management des données de la recherche, gestion de contenu et documentation des données - 2017 Vandoeuvre-lès-Nancy

7.5.1 Vers des entrepôts de données de confiance ou certifiés

Dans le but de pouvoir être pérennisées et réutilisées, les données ont intérêt à être déposées dans des entrepôts. Déposer des données dans des entrepôts nécessite un certain nombre de prérequis pour assurer la qualité des données déposées :

- favoriser le dépôt des données dans des formats ouverts interopérables,
- avoir des données validées et présentant un code renseignant sur la qualité des données,
- avoir des métadonnées descriptives bien renseignées et faisant partie d'un thesaurus identifié.

Il est également nécessaire de se préoccuper de la qualité des entrepôts que l'on va choisir pour y déposer les données. Pour être dignes de confiance, les entrepôts doivent également répondre à certains prérequis et spécifications qui, si besoin, peuvent amener à une certification.

Dans le cadre du [séminaire du réseau SIST20](#), Aude Chambodut a présenté les fonctionnalités « TRUST » qui permettent d'avoir confiance dans un entrepôt, et en quoi consiste l'intérêt d'une certification « Core Trust Seal ». Comme [Le Plan national pour la Science ouverte](#), elle nous rappelle que : « rendre les données FAIR tout en les préservant sur le long terme nécessite d'avoir des entrepôts fiables, dotés d'une gouvernance et de cadres organisationnels durables, d'une infrastructure fiable et des politiques globales soutenant des pratiques approuvées par la communauté ».

« Pourquoi et comment aller vers la certification Core Trust Seal ? »

[vidéo](#) : Aude Chambodut Séminaires SIST20

CoreTrustSeal est un organisme communautaire sans but lucratif qui promeut le développement d'infrastructures de données durables et fiables et spécifie les critères de conformité qui permettent de certifier un entrepôt.

La [Research Data Alliance](#) recommande les [critères de conformité de Core trust Seal](#), qui spécifient un entrepôt de confiance.

S'ils ne sont pas certifiés, les entrepôts de confiance devraient, a minima, respecter les 5 principes TRUST : transparence (Transparency), responsabilité (Responsibility), orientation vers l'utilisateur (User focus), durabilité (Sustainability) et technologie (Technology).

- Transparence : La transparence signifie que la gestion de l'entrepôt doit être vérifiable par des preuves accessibles au public.
- Responsabilité : La responsabilité implique de fournir toutes les garanties d'intégrité des données, de fiabilité et de pérennité de l'entrepôt.
- Orienté utilisateur : implique de veiller aux attentes des utilisateurs en matière de dépôt de données.
- Durabilité : demande à ce que les collections de données soient préservées sur le long terme.
- Technologie : implique de fournir l'infrastructure et les capacités nécessaires pour obtenir des services sécurisés, pérennes et fiables.

Les principes TRUST donnent aux utilisateurs l'assurance qu'ils bénéficient d'entrepôts sûrs avec des moyens durables.

7.5.2 Entrepôts en SHS

En sciences humaines et sociales, NAKALA est un service proposé par l'infrastructure de Recherches « Huma-Num » pour déposer, documenter et diffuser les données de la recherche. Il permet de rendre les données interopérables et de diffuser très simplement, dans des publications électroniques, les données déposées dans NAKALA.

L'entrepôt de données de recherche NAKALA, est destiné à accueillir, conserver et rendre visibles et accessibles les données de recherche selon les principes FAIR. Il permet d'enregistrer des données numériques de tout type (fichiers texte, son, images, vidéo), de les décrire en vue de les exposer et les rendre réutilisables et citables. Ainsi le dépôt de données dans NAKALA va offrir des services sur plusieurs étapes du cycle de vie de vos données, sur la préservation, la publication et la réutilisation. Le service NAKALA offre deux niveaux de préservation :

- Un niveau par défaut qui est mis en pratique dès lors qu'une donnée est enregistrée dans NAKALA. La donnée est décrite, contextualisée et stockée de manière sécurisée. Au titre de la préservation, déposer et décrire ses données dans NAKALA apporte la garantie d'une conservation des données dans un environnement sécurisé. Accompagnée d'une description, elle apporte aussi une conservation au niveau intellectuel garantissant sa compréhension à long terme.
- Les données peuvent être organisées et regroupées dans des collections qui elles mêmes peuvent être décrites et identifiables. Le projet NAKALA_Press permet de présenter de façon personnalisable vos collections en complément des pages de recherche et de consultation disponibles directement dans NAKALA.

On trouvera ci dessous les présentations nécessaires pour utiliser l'entrepôt nakala :

- [Un tutoriel pour déposer et documenter ses données dans nakala](#)
- [Utilisation de nakala pour déposer et diffuser les données de la recherche](#)
- [Pour découvrir pas à pas l'interface de NAKALA ainsi que les modalités de l'authentification](#)

7.5.3 Déposer/Publier dans des entrepôts institutionnels

Déposer dans des Entrepôts.. lesquels ? comment ?

Il existe beaucoup d'entrepôts de données, de nature et de qualité différentes. Certains sont des entrepôts Institutionnels (Portail Data INRAE, DataSuds, Didomena, ...), d'autres sont thématiques (PANGAEA pour les données environnementales, SEANOE spécifique aux données marines ...) ou généralistes (Dryad, Zenodo, ...). Pour aider à trouver et à choisir un entrepôt, des catalogues sont disponibles : <https://cat.opidor.fr/>, <https://www.re3data.org/> et <https://fairsharing.org/databases/>. Des entrepôts spécifiques peuvent être suggérés (ou imposés) par le journal dans lequel on dépose un article, mais aussi par le financeur, le consortium du projet ou l'institution dans laquelle on travaille. Il est conseillé de vérifier si l'établissement dans lequel on travaille a mis en place une politique de partage de données et de s'y référer pour éviter la dispersion des données tous azimuts.

7.5.4 Les Infrastructures de Recherche nationales

Le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation tient à jour la liste officielle des Infrastructures de Recherche nationales.

7.6 Publier un « Datapaper » pour valoriser et expliciter les données

Le data paper est un article scientifique sur les données : il permet de décrire un jeu de données de recherche (data, dataset), à l'aide d'informations plus précises et détaillées que celles qu'on peut trouver dans un « plan de gestion de données » (DMP), notamment en insistant sur :

- Les aspects méthodologiques,
- la qualité des données et de leur méthode de collecte et de traitement,
- l'originalité et la portée de ce jeu de données, ainsi que leur potentiel pour des utilisations futures (arguments décisifs pour l'acceptation de la publication),

- l'accès au jeu de données, dans un fichier attaché ou par un lien pérenne (URL, DOI) vers un entrepôt où le jeu est déposé et accessible.

On ne confondra pas les informations fournies par un data paper, avec celles qu'on l'on donne lorsqu'on rédige un Plan de Gestion de données. Ce sont des informations différentes [nous en parlons dans une FAQ](#)

Publier un data paper permet de :

- valoriser les données,
- faciliter leur réutilisation,
- leur apporter de la visibilité,
- les rendre plus facilement repérables et citables ; le data paper étant une publication citable, au même titre que tout article scientifique, il met en valeur ses auteurs en tant que créateur de données et permet la traçabilité des citations et des réutilisations.

Un data paper est une publication scientifique. Comme un article scientifique, il est validé par des « reviewer ». Mais c'est un article scientifique qui se distingue des articles présentant des résultats de recherche en plusieurs points :

- Il est centré sur un jeu de données et il a pour finalité de le décrire.
- Il se distingue d'un article scientifique traditionnel par le fait qu'il ne comporte pas d'hypothèses, ni d'interprétation, ni de discussion de résultats, ni de conclusions sur une question de recherche scientifiques

Selon les journaux et les communautés, la portée du « peer review » va varier. Certains vérifient uniquement la cohérence et la qualité de la description du jeux de données, d'autres évaluent les données elles-mêmes. Il est donc important de prendre en compte les politiques des revues par rapport aux données ¹.

La différence est plus délicate lorsqu'il s'agit de comparer un data paper et un article contenant des « *supplementary data* » qu'il décrit. Là en effet, les distinctions ne sont pas toujours claires, surtout du fait du manque de recul car ce sont des pratiques récentes et toujours émergentes. Certains data paper sont très très bref et ne vont pas beaucoup plus loin que ce qu'on trouve dans la fiche accompagnant le jeu de données dans l'entrepôt, d'autres sont beaucoup plus complexes et jouent plus profondément la carte de la réutilisation en tentant d'explicitier les implications des jeux de données et des traitements subis.

Le data paper est publié, en libre accès, sous la forme d'un article examiné par les pairs dans une revue scientifique classique publiant différentes formes d'articles, dont des data papers, ou dans un data journal, c'est-à-dire une revue contenant exclusivement des data papers.

Il n'existe pas, à ce jour, de *catalogues ou de répertoires* à proprement parlé, mais nous recommandons la consultation de ces listes de liens génériques :

- [CIRAD](#),
- [Forschungsdaten](#),
- [Datashare](#) et dans le domaine de la
- [bio-diversité](#).

Après avoir expliqué pourquoi le GBIF et l'éditeur de revues PENSOFT ont proposé le concept de data paper, Sophie Pamerlon en explique les avantages et comment les outils du GBIF facilitent la rédaction d'un datapaper en biodiversité à travers quelques exemples concrets.

Data papers : Une incitation à la publication de données sur la biodiversité,

Sophie Pamerlon :Système mondial d'information sur la biodiversité - GBIF Global Biodiversity Information Facility

On trouvera de nombreuses informations sur la création et l'évaluation de Data papers dans le Webinaire intitulé « DataPaper : une incitation à la qualification et à la réutilisation des jeux de données » organisé par l'«[Atelier Données](#)» du groupe de travail Données inter-réseaux de la MITI.

Sophie Pamerlon présente les avantages de publier un datapaper, ainsi que deux outils de rédaction de datapaper :

- IPT (Integrated Publishing Toolkit) qui facilite le remplissage des métadonnées et la production automatisée d'un manuscrit de Data Paper

1. <https://datascience.codata.org/articles/10.5334/dsj-2020-005/>

- ARPHA : Outil de rédaction qui facilite la mise en page, la soumission, le processus de relecture, la publication, l'hébergement et l'archivage d'articles scientifiques.

Exemple d'intégration du data paper à un workflow de publication de jeux de données : l'outil intégré de publication (IPT) du GBIF/ Retour d'expérience d'un producteur de data paper

Vidéo : Sophie Pamerlon, GBIF France – USM Patrimoine naturel

Pour se faire une idée d'un exemple de datapaper, Annegret Nicolai nous présente un [exemple de datapaper du projet bioBlitz](#) et les avantages et inconvénients qu'elle y trouve :

BioBlitz 2017 à la Station Biologique de Paimpont – un data paper de science citoyenne

Vidéo : Annegret Nicolai, (Univ. Rennes 1 – UMR ECOBIO, Station Biologique de Paimpont)

Dans sa présentation, Clémentine Cottineau nous indique quels sont les principes et le processus d'évaluation d'un datapaper pour la revue *Cybergeog* : Retour d'expérience et difficultés rencontrées. On trouvera sur *cybergeog* un [exemple de recommandations aux auteurs pour un datapaper](#).

Évaluer un data paper : retour d'expérience de la revue *Cybergeog*

Vidéo : Clémentine Cottineau, CNRS – Centre Maurice Halbwachs, Denise Pumain, Univ. Paris 1 – UMR Géographie-Cités, Christine Kosmopoulos, CNRS – UMR Géographie-Cités

Victor Gay nous présente un retour d'expérience de rédaction d'un [datapaper publié sur HAL](#) selon le modèle de la revue *Scientific Data*. Il nous présente la production d'un data paper du point de vue d'un chercheur. Après avoir exposé sa recherche et les données produites, il explique pourquoi il a décidé de rédiger un data paper, la manière dont il s'y est pris pour le dépôt des données et la rédaction, avant de revenir sur les choix de dissémination et le rôle des métiers de l'accompagnement de la recherche dans l'ensemble du processus.

Retour d'expérience d'un producteur de data paper

Vidéo : Victor Gay, Univ. Toulouse 1 – École d'Économie de Toulouse

Joachim Schöpfel, propose une synthèse des différentes communications du webinaire en indiquant qu'un data paper fournit l'information « on the *what, where, why, how and who of the data* ». Il revient ainsi sur l'intégration des data papers dans les pratiques des communautés, leur diversité de forme, leurs objectifs, leur évaluation, leur impact, le rôle des différents métiers de la recherche dans leur production, pour finir sur les perspectives en la matière et ouvrir ainsi sur des échanges avec les participants.

Synthèse du webinaire et échanges

Vidéo : Joachim Schöpfel, Université Lille 3 – GERiCO

En guise d'exercice de conclusion, Wilfried Heintz nous fait part de sa conception d'une gestion pérenne des données scientifiques, en reliant nos différentes actions depuis l'étape initiale de la rédaction d'un DMP (Plan de gestion des données) jusqu'à la publication d'un DataPaper :

Du Plan de Gestion des Données au Datapaper : suivi des données scientifiques tout au long de leur cycle de vie

Wilfried Heintz, UMR 1201 Dynafor [Séminaire SIST18 Observatoire Versailles](#)

Gestion pérenne des données scientifiques : du plan de gestion de données au datapaper.

Wilfried Heintz, UMR 1201 DynaforStorage Day 2018, Paris.

7.7 Publier des données grâce au web des données et au web sémantique

Selon Wikipedia, « le Web sémantique est une extension du Web standardisée par le World Wide Web Consortium (W3C). Ces standards encouragent l'utilisation de formats de données et de protocoles d'échange normés sur le Web, en s'appuyant sur le modèle Resource Description Framework (RDF). Le Web sémantique est par certains qualifié de web 3.0.»

Selon Wikipedia, « le Web des données (linked data, en anglais) est une initiative du W3C (Consortium World Wide Web) visant à favoriser la publication de données structurées sur le Web, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un réseau global d'informations.»

Tim Berners-Lee (inventeur du Web et directeur du W3C), qui supervise le développement des technologies communes du Web sémantique a défini le web sémantique comme « une toile de données, données qui peuvent être traitées directement et indirectement par des machines pour aider leurs utilisateurs à créer de nouvelles connaissances ».

Rendre nos Données accessibles et interopérables sur le Web,

Franck Michel (I3S - UMR 7271, CNRS - Univ. Nice Sophia), 2015 mots-clés : SPARQL, web sémantique, RDF, SKOS, OWLAction nationale de formation RBDD 2015

Rendre les données interopérables sur le web est le sujet essentiel de cette présentation très complète. Après avoir posé le contexte, Franck Michel développe le sujet en déroulant le plan suivant :

- The Web of Data and the Semantic Web
- Create, reuse and link vocabularies
- Populate the Web of Data
- Publish Linked Open Data on the Web

Il détaille le modèle RDF (Resource Description Framework) du W3C, puis, le langage de requêtes SPARQL. Il explique ensuite le standard SKOS (Simple Knowledge Organization System) utilisé pour représenter les vocabulaires contrôlés, les taxonomies et thesauri. Il termine en montrant comment publier des données ouvertes sur le web.

Atelier « Mise en place d'un SPARQL EndPoint. Servir du RDF via HTTP avec Jena et Fuseki »

Wilfried Heintz (Unité Mixte de Recherche « Dynamiques et écologie des paysages agriforestiers »), 2015

Cet atelier technique est organisé selon le plan suivant :

- Présentation de l'outil Jena
- Prérequis et préparation du serveur
- RDFizer les métadonnées ou les données
- Installation de Fuseki
- Exemples d'exploitation du Sparql Endpoint

Tous les éléments exposés dans ce chapitre sont nécessaires et importants pour mettre en place une bonne publication et diffusion des données de la science.

Conclusion

La rédaction de ce guide a été motivée d'une part, par les évolutions récentes liées aux problématiques de gestion des données de la recherche dans le cadre d'une science ouverte, et d'autre part, par le regroupement et la réflexion interdisciplinaire des membres de réseaux de la MITI et d'Intituts CNRS.

Les réseaux métiers sont en effet particulièrement actifs et investis dans la veille technologique et la diffusion de savoirs nécessaires pour une bonne gestion FAIR des données. Grâce à leurs actions ils constituent le relais nécessaire pour diffuser les bonnes pratiques utiles pour le travail dans les laboratoires. En ce sens à travers les multiples séminaires et formations et ANF qui ont été organisés, on peut dire qu'ils sont vecteurs de bonnes pratiques et de diffusion du savoir dans la gestion des données de la recherche.

Eu égard au travail de soutien important dans la gestion des données, le groupe « Atelier Données » créé via la MITI a estimé intéressant et nécessaire de rendre compte des compétences et expertises présentes dans nos réseaux, sous la forme d'un guide de bonnes pratiques inhérentes aux données de la recherche.

Du fait des différentes approches, outils, concepts et vocabulaires entre nos différents métiers, nous avons retenu la solution fédératrice de relier nos actions aux étapes communément admises du « cycle de vie des données » auquel nous avons également ajouté une étape « Imaginer et Préparer » et « concevoir planifier », en les distinguant, pour bien prendre en compte les phases préparatoires de planification d'un projet. L'intérêt de cette représentation est de bien montrer toutes les étapes nécessaires pour aboutir à une publication des données, et les rendre réutilisables. Ainsi les données de la recherche pourront être mises à profit dans d'autres projets.

Le cycle de vie doit assurer aux données les meilleures conditions pour leur utilisation, leur archivage pérenne, et leur réutilisation pour d'autres besoins et d'autres projets que ceux pour lesquels elles ont été initialement constituées.

Outre des différences méthodologiques liées aux disciplines qu'on retrouvera dans les étapes 3 (collecter), 4 (traiter), 5 (analyser), les réseaux se retrouvent sur des concepts communs dès lorsqu'il faut préparer un projet (étapes 1,2), préserver les données (étape 6) ou publier et diffuser les données (étape 7).

L'originalité de ce document est qu'il rassemble la majeure partie de la production des réseaux métiers de ces dernières années, dans le cadre de la gestion des données de la recherche.

Ce document est donc un condensé des actions menées autour de la gestion des données de la recherche. Il est le fruit d'un travail collaboratif qui a consisté à collecter, sélectionner et mettre à disposition des ressources vers les actions phares des réseaux métiers, enrichis d'informations et de conseils.

Ce guide n'a pas la prétention d'être exhaustif, mais il illustre les thèmes de fort intérêt de ces dernières années menés par les réseaux métiers qui s'inscrivent dans la politique nationale liée à la science ouverte. Il sera complété au fil du temps par d'autres thèmes et actions d'intérêt organisés par les réseaux.

Les pratiques et conseils cités dans ce guide ne se substituent pas aux recommandations présentées par les agences de financement, les établissements, ou les instituts, . . . , mais sont là pour éclairer, accompagner les personnels de la recherche en charge de la gestion des données.

Il est clair désormais qu'il faut considérer la gestion de données comme une tâche à part entière dans les projets de recherche. On doit désormais anticiper comment les données seront acquises/collectées, stockées, diffusées et bien entendu respecter les règlements de manière à « ouvrir les données autant que possible, les fermer autant que nécessaire.

Le lecteur aura par ailleurs connaissance des infrastructures existantes. Il pourra se positionner sur les pratiques utilisées en fonction de sa discipline, et enfin faire appel et se rapprocher des réseaux métiers pour l'aider à une bonne gestion des données.

Nous espérons à travers ce guide apporter notre pierre à l'édifice pour une meilleure prise en compte du travail, consacré aux données de la recherche, pour qu'elles puissent être accessibles, bien documentées, réutilisables et donc réutilisées dans le cadre de la science ouverte.

Archiver : l'archivage consiste à ranger un document dans un lieu où il sera conservé pendant une période plus ou moins longue et d'y associer les moyens pour réutiliser les données : la réutilisation se faisant en ajoutant de l'intelligence à la sauvegarde. Le contenu des documents archivés n'est pas modifiable. Par contre le contenant (format) des documents archivés peut être modifié (pour éviter l'obsolescence logicielle).

DMP (PGD) : un « DMP » est un « Data Management Plan ». Sa traduction française est « PGD » pour « Plan de gestion de données ». Un Plan de Gestion de données est un document formel, établi en début de projet qui formalise la manière dont seront gérées les données à différents points de vue : stockage, permissions d'accès, sécurité etc.

Pérenniser : ce verbe est souvent utilisé à la place de préserver quand on pense archivage pérenne. L'archivage pérenne a pour fonction d'assurer la conservation à long terme des données, leur accessibilité tout en préservant leur intelligibilité, comme rendre accessible en lecture des données immuables (archives de documents administratifs, données de mesures expérimentales, résultats de simulations coûteuses à produire, etc.).

Préserver : cette action fait référence au fait de garantir, protéger, mettre à l'abri, sauver d'un dommage ou d'une destruction (cf. notion de sauvegarde) et au fait de tenir dans le même état, en bon état (intelligible). Elle fait aussi référence à la notion de permanence dans le temps (cf. notion d'archivage). Le synonyme « conserver » est utilisé quand il est fait référence à une politique.

Sauvegarder : la sauvegarde consiste à dupliquer les données sur un support numérique externe à celui où elles sont stockées. L'objectif est de pouvoir les retrouver en cas de perte ou de dégradation de l'organe de stockage. Il s'agit d'une sauvegarde octet par octet dans une perspective de court ou de moyen terme. La recherche de la préservation de l'intelligibilité des données n'est pas un élément pris en compte.

Stocker : c'est l'étape première qui consiste à déposer les données sur un support numérique pour les rendre accessibles. Cela peut être un ordinateur personnel, un disque partagé ou tout autre organe de dépôt. Le stockage permet d'assurer la continuité de l'exploitation sur du court terme. A ce stade, la donnée n'est ni sauvegardée et ni sécurisée.

Le paysage des infrastructures destinées à la recherche scientifique est vaste. Cette section est destinée à la présentation d'infrastructures européennes, nationales, thématiques dans différents domaines. Cette liste n'est bien sûr pas exhaustive et sa caractéristique principale est d'être alimentée par les présentations qui ont été réalisées au cours des événements organisés par les réseaux métiers. Ces exposés ciblent donc en général le public de ces réseaux.

10.1 Infrastructures Européennes

Ces infrastructures font régulièrement l'objet de présentations qui permettent de comprendre leur organisation, leur mode de fonctionnement, et de suivre leurs évolutions.

10.1.1 European Open Science Cloud (EOSC), Infrastructure Européenne pour la science ouverte (en cours de montage)

Le cloud européen EOSC a été abordé par Volker Beckman, chargé de mission CNRS-EOSC et Directeur adjoint scientifique Calcul et Données IN2P3/CNRS, lors des [JCAD 2019](#). Il explique comment concrètement les chercheurs peuvent utiliser EOSC, et fait le point sur les principales questions que vous vous posez peut-être sur EOSC : Qu'est ce que EOSC ? Où en est-on et quelles sont les prochaines étapes ? Comment puis-je contribuer et/ou bénéficier ? EOSC en France. Il présente la stratégie européenne d'élaboration de ce Cloud européen lancé en 2016, qui offre aux chercheurs des services, l'accès à des données et à d'autres ressources fournies par des infrastructures de recherche publiques nationales, régionales et institutionnelles en Europe. Partant des nombreux appels à projets qui ont été lancés pour construire ce cloud (EOSCpilot, EOSC-hub, EOSC-Pillar, etc.), il expose les possibilités de collaboration. Depuis 2019, une structuration se met en place : une gouvernance temporaire et une implication forte des ministères nationaux comme têtes de file des communautés scientifiques impliquées. Ces travaux préparatoires devraient déboucher sur d'importantes opportunités de financement dans le « Horizon Europe Work Programme ». Le lien avec l'infrastructure nationale France Grilles est également décrit.

EOSC a fait l'objet d'une [journée spéciale au CNRS en janvier 2020](#). Cette journée a en particulier permis de faire le point sur l'implication française dans le projet à travers plusieurs exemples. Une journée nationale [EOSC-France](#) aura lieu en février 2021.

Plus récemment dans le cadre de l'atelier Dialogu'IST, Réseau Renatis, le 9 juillet 2020, une autre présentation de Volker Beckman a fait le point sur l'"European Open Science Cloud (EOSC), opportunités pour la recherche en France.

Enfin, lors des JCAD 2020, Volker Beckmann a présenté dans « EOSC en France: défis et opportunités. » (vidéo) les prochaines étapes prévues.

EOSC en France : défis et opportunités.

Vidéo : Volker Beckmann, Chargé de mission EOSC France pour le ministère.JCAD 2020

EOSC est ouvert à la participation de fournisseurs de services, que ce soit des services de stockage ou de traitement de données ou encore des services de publication. Différents projets européens successifs ont pour objectif de préparer EOSC. Dans le cadre du projet **EOSC-Pillar**, une enquête internationale a été menée dans cinq pays (Allemagne, Autriche, Belgique, France et Italie) pour connaître l'état des Infrastructures de Recherche, e-infrastructures, universités et organismes financeurs par rapport à une éventuelle participation à EOSC. Un aperçu des [résultats de cette enquête](#) a été présenté lors des JCAD 2020.

Que nous apprend l'enquête du projet européen EOSC-Pillar ?

Vidéo : Geneviève Romier, Centre de Calcul de l'IN2P3, CNRSJCAD 2020

10.1.2 Infrastructure Européenne EGI (infrastructure proposant différents services basés sur des infrastructures grille et cloud), egi.eu

L'infrastructure EGI a été présentée de façon complète en 2018 lors des JCAD, Journées Calcul et Données : la fédération, les participants, le catalogue de services, les utilisateurs, le positionnement dans EOSC. Le projet EOSC-Hub qui participe à la construction d'EOSC est également détaillé. C'est la présentation à consulter si vous souhaitez savoir ce qu'est EGI.

EGI, the EOSC and the Hub

Vidéo [EGI, the EOSC and the Hub](#) Yannick Legré, directeur de la fondation EGIJCAD 2018

— Pour compléter, vous pouvez aussi consulter [une présentation plus détaillée de son mode de fonctionnement](#)

Une présentation des évolutions a eu lieu en 2019 avec un focus sur quelques services EGI - Check-in, Cloud, Stockage, Notebooks, Application on Demand - et quelques communautés utilisatrices.

Update about EGI and services to support user communities and Federation's participants

Vidéo : Baptiste Grenier, fondation EGIJCAD 2019

10.2 Infrastructures Nationales

CAT OPIDoR recense les services dédiés aux données de la recherche en perspective du cycle de vie des données et par type de service. Ce catalogue se présente sous la forme d'un wiki, n'hésitez donc pas à le compléter si besoin.

Les présentations détaillées de différentes infrastructures et de leurs services citées ci-dessous peuvent vous permettre à la fois de découvrir le large paysage de l'offre en Europe et en France mais aussi de vérifier rapidement si les critères de votre projet peuvent être remplis par les différentes infrastructures et leurs offres de services.

10.2.1 Infrastructures de travail collaboratif

Dans la phase de montage de projet, il convient de choisir et de mettre en place des outils de gestion de projets tels que :

- des listes de discussion fournies par un service de gestion de listes,
- des outils de partage de documents et de données dans des dossiers partagés en réseau ou de type « service de cloud »,
- une plate-forme de gestion de projet de type « redmine » ou autre.

Pour cela il est utile de connaître les possibilités et ressources internes à l'unité et celles fournies par l'institution ou des partenaires extérieurs : université, CNRS, Renater, etc. . .

10.2.2 Infrastructures de traitement de données, calcul, stockage

De nombreuses infrastructures offrent des services à la communauté scientifique. Il convient de choisir celles qui conviennent le mieux pour chaque projet.

Infrastructure France Grilles, www.france-grilles.fr

Le catalogue de services de France Grilles propose des services de traitement de données qui s'appuient sur une infrastructure de grille et une infrastructure cloud ainsi que des services de stockage de données. L'ensemble est interconnecté permettant aux données stockées d'être traitées grâce aux services grille et cloud.

Un poster présente le service de stockage de France Grilles :

FG-iRODS : un service de gestion de données pour les communautés scientifiques à l'échelle nationale et européenne basé une infrastructure fédérée

Emmanuel Medernach, Jérôme Pansanel, Raphaël Flores, Christine Gondrand, Patrick Moreau, Vincent Negre, Genevieve RomierJCAD 2019

Un autre poster a pour objet le cloud France Grilles : FG-Cloud : the French Academic Cloud for Scientific computing

Groupe FG-CloudJCAD 2018

Il est aussi important de se faire une idée à travers des retours d'expérience réalisés par des collègues. En voici quelques exemples récents :

Le projet Phénome, Infrastructure nationale de phénologie végétale, regroupe sur neuf sites des plateformes expérimentales de phénotypage haut-débit (champ, serre, omique). Un système complet a été mis en place pour ce projet, système qui s'appuie sur les services FG-iRODS et FG-Cloud.

Déploiement de la plateforme de traitement des données phénotypage haut débit 4P sur l'infrastructure France Grilles

Vidéo : Vincent Negre, Eric David, Philippe Burger, Romain Chapuis, Boris Adam, Anne Tireau, Patrick Moreau, Antony Tong, Samuel Thomas, Gallian Colombeau, Pascal Neveu, Jérôme Pansanel, Frédéric Baret, Marie Weiss JCAD 2019

Centres de calcul de GENCI

Les plateformes de calcul intensif nationales, ainsi que leurs évolutions, sont régulièrement présentées par l'opérateur GENCI (Grand Equipement National pour le Calcul Intensif). GENCI et ses trois centres nationaux fournissent des moyens de calcul de niveau « Tier 1 » pour les utilisateurs nationaux :

- Ainsi, lors des JCAD 2019 a été abordé le supercalculateur Jean Zay dont une partie est dédiée à l'Intelligence Artificielle (IA).
 - La [présentation de GENCI aux JCAD 2018](#) propose une approche plus générale et introduit également l'écosystème national des mésocentres, et les projets européens liés au calcul intensif.
 - Une [présentation plus récente de GENCI aux JCAD 2020](#) présente aussi des résultats obtenus sur le supercalculateur Jean Zay.
-

Actualité GENCI

vidéo Philippe Lavocat représenté par Elise Quentel et Jean-Philippe Proux, GENCI JCAD 2019

GENCI, une TGIR active au niveau régional, national et européen.

Vidéo : Stéphane Requena, Directeur innovation et technologie, GENCI JCAD 2020

Infrastructures des mésocentres et centres régionaux, quelques exemples

Au niveau régional et local, les mésocentres de calcul fournissent des ressources et un accompagnement de proximité plus souple et en général plus facile d'accès que les ressources nationales.

Une [présentation réalisée en 2017](#) fait un retour sur les mésocentres au cours des dix dernières années.

Les mésocentres ont fait l'objet de présentations régulières, à la fois techniques et organisationnelles, lors des journées mésocentres organisées jusqu'en 2017 par le réseau Calcul :

- [Journée mésocentres 2016](#)
- [Journée mésocentres 2017](#)

De nombreux mésocentres ont participé à l'équipex Equip@Meso coordonné par GENCI et qui a fait l'objet par exemple, lors des JCAD 2018, d'[un poster](#) d'Elise Quentel de GENCI, qui synthétise les informations

Lors des JCAD 2019, Cyrille Toulet a présenté l'intégration du cloud OpenStack du mésocentre de Lille dans plusieurs fédérations de cloud nationales et internationale. Il explique l'intérêt de ces intégrations et les aspects techniques et donne des cas d'utilisation dans différentes disciplines.

Intégration d'un cloud OpenStack à plusieurs fédérations de cloud

Vidéo : Cyrille TOULET, Mésocentre de Lille, Université de Lille JCAD 2019

Lors des JCAD 2018, Jérôme Pansanel a présenté la plate-forme SCIGNE de l'Institut Pluridisciplinaire Hubert Curien de Strasbourg. Cette plate-forme est accessible aux utilisateurs régionaux, nationaux et européens.

La plateforme SCIGNE : présentation et utilisation du service de Cloud Computing

Vidéo : Jérôme Pansanel, Institut Pluridisciplinaire Hubert Curien, CNRS JCAD 2018

La présentation d'autres centres est disponible sous forme de poster, comme par exemple le [Pôle Scientifique de Modélisation Numérique \(PSMN\)](#) de l'ENS de Lyon, le [Centre Blaise Pascal](#) de l'ENS de Lyon et le [mésocentre CALMIP](#) à Toulouse.

Qu'est-ce que le PSMN de l'ENS de Lyon ?

Coraline Petit, Cerasela Iliana Calugaru, Micaël Calvas et Loïs Taulelle, Pôle Scientifique de Modélisation Numérique, École Normale Supérieure de Lyon JCAD 2018

Le Centre Blaise Pascal : de l'hôtel à projets au centre d'essais.

Emmanuel Quemener et Micaël Calvas, Centre Blaise Pascal, École Normale Supérieure de Lyon JCAD 2018

Mésocentre CALMIP

Mickaël Duval, Nadine Marouzé, UMS 3667 CALMIP - Université de Toulouse, INPT, Université Paul Sabatier, INSAT, ISAE-SUPAERO et CNRS JCAD 2018

Il est aussi possible d'utiliser des ressources fournies par plusieurs mésocentres dans le cadre d'un même projet. Ce retour d'expérience dans le domaine de la chimie en est un témoignage.

Calculs de chimie quantique distribués entre méso-centres avec Quantum Package

Vidéo : Anthony Scemama, Patrick BOUSQUET-MELOU, Marie-Sophie Cabot, Nicolas Renon JCAD 2019

Infrastructure pour les expériences à grande échelle en informatique

SILECS, « Super Infrastructure for Large-scale Experimental Computer Science », est dédiée aux expériences à grande échelle en informatique basée sur les infrastructures FIT et GRID'5000. Cette infrastructure, tout en conservant les objectifs de FIT et GRID'5000, vise de nouveaux challenges : Internet des objets - Internet of Things (IoT) et Clouds, nouvelles générations de plateformes Cloud et de piles logicielles (Edge, FOG), applications de Data streaming, gestion de volumes de données importants, mobilité...

SILECS, une infrastructure pour les expériences à grande échelle en informatique

Vidéo : Frédéric Desprez, Christian Perez, Inria JCAD 2019

Dans la présentation « Slices, towards a Scientific Large-Scale Infrastructure for Computing - Communication Experimental Studies », aux JCAD 2020, est détaillée la proposition d'infrastructure européenne Slices, « Super Infrastructure for Large-Scale Experimental Computer Science », dont l'objectif est de construire sur 15 pays une infrastructure dont SILECS est la partie française.

Slices, towards a Scientific Large-Scale Infrastructure for Computing - Communication Experimental Studies.

Vidéo : Christian Perez, Inria JCAD 2020

10.2.3 Infrastructures régionales de gestion de données

Des initiatives régionales proposent des infrastructures de gestion et valorisation des données pour la recherche.

A Toulouse, le mésocentre CALMIP étoffe son offre de services et construit CALLISTO, une interface pour le partage et l'analyse semi-automatique de données. CALLISTO propose ainsi une aide à la rédaction de Plans de Gestion de Données sur les aspects techniques, une plateforme de partage de données proche des utilisateurs et en lien avec les ressources du supercalculateur pour permettre la réutilisation des données hébergées.

Une présentation du projet a été réalisée lors des JCAD 2019, puis une démonstration a eu lieu lors des JCAD 2020 qui permettent de voir les évolutions.

CALLISTO : Une interface pour le partage et l'analyse semi-automatique de données.

Vidéo : Thierry Louge - du mésocentre CALMIP, Toulouse. JCAD 2019

Partage et analyse semi-automatique de données pour un mésocentre de calcul : fonctionnalités et avancement de CALLISTO pour CALMIP. Suivi de la démonstration de CALLISTO à CALMIP.

Vidéo : Thierry Louge - du mésocentre CALMIP, Toulouse. JCAD 2020

En Bourgogne Franche-Comté, le projet dat@UBFC a pour objectif la création d'un service de gestion des données de la recherche pour la communauté scientifique de l'Université de Bourgogne Franche-Comté. Ce projet fait le lien entre dat@OSU (Description et référencement des données de recherche de l'OSU THETA) et le datacenter régional UBFC. Il a fait l'objet d'une présentation détaillée aux JCAD 2020.

Projet dat@UBFC : création d'un service de gestion des données de la recherche pour l'Université de Bourgogne Franche-Comté.

Vidéo : Sylvie Damy, laboratoire Chrono-Environnement, Besançon. JCAD 2020

10.2.4 Les « datalakes » ou « lacs de données »

De nouvelles infrastructures de stockage de grandes quantités de données apparaissent dans le paysage. Plusieurs présentations vous permettront de vous faire une idée sur ce qu'est un Data Lake même si le concept n'est pas encore complètement défini et figé.

Jean-Pierre Gleyze présente le contexte du CNES, quelques applications et leur dimensionnement. Les points de vue des utilisateurs et des administrateurs des infrastructures sont exposés, aussi cet exposé intéressera autant les informaticiens que les utilisateurs scientifiques.

Keynote : Infrastructures de traitement de données : vers un datalake CNES

Vidéo : Jean-Pierre Gleyze, CNES JCAD 2019

CEBA, a pour ambition la création d'un « grand » observatoire de l'environnement en Auvergne, unique en Europe. Ce cloud environnemental permettra la gestion des données à tous les stades de leur cycle de vie. Il propose différents services comme un site Web, des outils d'ingestion, un moteur d'indexation, un catalogue de données, des outils de visualisation qui s'appuient sur une infrastructure incluant des bases de données et système de fichiers.

CEBA, un data lake dédié à l'observation des écosystèmes environnementaux

Vidéo : Francis Ogereau, Vincent Breton, Alexandre Claude, David Grimbichler, Antoine Mahul, Gilles Mailhot, Jérémy Mezhoud, Christine Plumejeaud, Laurent Royer, Estelle Théveniaud, Richard Vandaele, David SarramiaJCAD 2019

Le projet international DOMA n'est pas un projet de DATALAKE ni un produit mais une organisation qui participera à définir ce que sera un DATALAKE. Ses objectifs sont de suivre les avancées et les développements, être un forum de partage d'informations et veiller à l'interopérabilité des différentes solutions de stockage. Une première présentation de DOMA a eu lieu en 2018 et une présentation des évolutions en 2019.

Projet DOMA : Réflexions et axes de recherche pour les services de stockage de données scientifiques à l'horizon 2025.

Vidéo : Eric Fede, Centre de Calcul de l'IN2P3, CNRSJCAD 2018

Projet DOMA : Retour sur la première année des études faites autour des futurs services de stockage de données scientifiques.

Vidéo : Eric Fede, CC-IN2P3, CNRSJCAD 2019

10.2.5 Infrastructures pour l'information scientifique et technique

L'Institut de l'Information Scientifique et Technique (INIST), unité de service du CNRS déploie ses activités vers un projet d'ingénierie des connaissances qui s'articule autour de 3 axes principaux : « **Analyse et fouille de l'information** », « **Valorisation des données de la recherche** », « **Accès à l'information scientifique** ». Claire François présente ici un **panorama des outils et services proposés aux chercheurs** : portail d'accès aux ressources électroniques (bibCNRS), plateforme d'accès aux archives scientifiques (ISTEX), une suite logicielle de mesure des usages des ressources électroniques (EzPAARSE.EzMEASURE) pour faciliter l'accès à l'information scientifique aux chercheurs. Elle présente également les outils de formation à distance et services d'accompagnement tels que Doranum, Conditor, CoRea pour optimiser le partage et l'interopérabilité des données de la recherche. Et pour finir les outils d'analyse et fouille de l'information scientifique tels que LOTERRE, ISTEX, LODEX ou VISA TM pour créer et gérer la terminologie scientifique et permettre le recueil des données sur les publications et la production d'indicateurs bibliométriques

Positionnement et offre globale de l'INIST dans le contexte IST en évolution

Claire FrançoisFredocs 2018 - Démarches innovantes en IST : expérimenter, proposer, (se) réinventer », 2018, Albi

10.2.6 Infrastructures pour les logiciels / les codes sources

Le projet **Software Heritage** a pour objectif de collecter, préserver et rendre disponible le code source (et son historique) de tous les logiciels publiquement disponibles. Cette présentation explique le contexte et les motivations qui ont donné lieu à Software Heritage, puis, l'architecture mise en oeuvre.

Software Heritage : source code archival and analysis at the scale of the world

vidéoStefano Zacchiroli, InriaJCAD 2019

10.2.7 Infrastructures thématiques

Des portails nationaux et européens organisés autour d'une thématique scientifique existent par ailleurs. Ils donnent accès à des pôles de données qui fédèrent dans différentes disciplines, des activités de gestion et valorisation des données.

Au sein de l'infrastructure de recherche « Data Terra » on trouve par exemple [Odatis](#) qui est un pôle de données et de services pour l'océan, [Theia](#), une infrastructure de données et de services dédiée aux données spatiales d'observation de la terre, [Aeris](#) qui s'intéresse aux données et services pour l'atmosphère mais aussi [Form@ter](#) ou [PNDB](#) (Pôle National de Données de Biodiversité).

Les projets de recherche au sein de ces infrastructures ont donné lieu à des retours d'expériences qui témoignent de spécificités disciplinaires dans la gestion des données de la recherche.

Sismologie et géodésie

Dans le domaine de la sismologie et de la géodésie, Résif (Réseau sismologique et géodésique français), une infrastructure de recherche pour l'observation des déformations de la terre a été mise en place. Catherine Pequegnat lors de son intervention aux Fredocs 2013, présente la spécificité des données de sismologie, les réseaux de capteurs associés à ces données et pointe les manques et les avancées en termes de description, de norme et de formats pour la transcription et l'échange des données. Elle aborde aussi la question des possibilités d'accès aux données. Elle présente ensuite la TGIR Résif, son projet, son centre de données national (possibilités d'hébergement et de stockage) et sa contribution au projet européen EPOS

« RESIF (données sismologiques et géodésiques) »

Catherine Pequegnat, Institut des Sciences de la Terre, Observatoire des Sciences de l'Univers de GrenobleFredocs2013 - Gestion et valorisation des données de la recherche, 2013, Aussois

Astronomie

Soizick Lesteven, retrace ensuite les missions du centre astronomique de Strasbourg qui depuis 1972 collecte des informations « utiles » sur les objets astronomiques. Elle détaille les processus de collecte et de traitements effectués pour mettre à disposition des données. Elle explique comment se crée l'interopérabilité des données entre les bases « Simbad », « Vizier », « Aladin » et leur interaction avec le dictionnaire des nomenclatures des objets célestes. Elle explique ainsi le processus de création de contenu de qualité, des données d'observations réutilisables et à forte valeur ajoutée pour les équipes de recherche.

La gestion des données astronomiques à l'Observatoire Astronomique de Strasbourg

Soizick Lesteven, Observatoire Astronomique de StrasbourgFredocs2013 - Gestion et valorisation des données de la recherche, 2013, Aussois

Sciences Sociales

Dans cette présentation très complète bien qu'ancienne, Roxane Silberman fait une description du Réseau Quetelet, composante de la TGIR Progedo qui est une banque française de données pour les sciences sociales. Elle indique la spécificité des données collectées au sein de ce réseau qui peuvent être des données individuelles, qualitatives, parfois de santé publique et les enjeux de protection de la vie privée associés à ces données. Elle retrace l'historique des premières banques de données pour lesquelles se posait déjà la question de l'accès et du partage des données à caractère individuel. Ce réseau s'inscrit dans une infrastructure européenne, le CESSDA (réseau européen des archives de données) et fournit des métadonnées accessibles à tous et en particulier aux chercheurs. Après avoir expliqué les grandes fonctions de Quetelet (dépôt, archivage, documentation, diffusion des données), elle revient sur quelques enjeux pour l'état, les politiques publiques et autres acteurs économiques, le paysage qui se dessine autour de la donnée et qui bouscule les frontières dans le contexte de l'open data.

Réseau Quetelet : Banques de données pour les sciences sociales

Roxane SilbermanQuetelet, 2013

Biodiversité

Dans le domaine de la biodiversité, le Pôle national de données de Biodiversité, PNDB, e-Infrastructure nationale de recherche est inscrite sur la feuille de route du MESRI depuis mars 2018. L'UMS PatriNat du Museum National d'Histoire Naturelle, MNHN, en est le maître d'oeuvre. Yvan Le Bras, lors des JCAD 2020 a présenté le PNDB et l'implémentation en cours.

Infrastructure PNDB, de la donnée de biodiversité au calcul scientifique via la métadonnée.

Yvan Le Bras – Laboratoire Patrinat du Museum national d'Histoire Naturelle, Paris.JCAD 2020

10.2.8 Plateformes d'archivage des données

Agrée par le Service Interministériel des Archives de France, le [CINES](#) (Centre Informatique National de l'Enseignement Supérieur) est le centre officiel d'archivage d'une partie de la production scientifique de nos établissements. Il offre une solution pour la conservation à long terme du patrimoine numérique (données scientifiques : issues d'observations ou de calculs, données patrimoniales, données administratives) et est impliqué le projet européen [EUDAT](#) visant à mettre en place une infrastructure européenne d'échange et de conservation de données.

A l'occasion d'une intervention au Gricad à Grenoble, Olivier Rouchon, détaille largement l'offre de service du CINES et témoigne de la variété des données archivées. Il retrace également le processus d'archivage et rend compte des défis et problématiques qu'il pose.

L'offre de service archivage du CINES

Oliver Rouchon, CINESArchivage numérique des données de la recherche, 2019, Grenoble [vidéo](#)

A noter : La TGIR Huma-Num propose une offre de service pour accompagner les producteurs de données tout au long du processus d'archivage à long terme vers un dépôt au CINES. (Voir la présentation de Michel Jacobson « [Archivage des données à Huma-Num](#) » présentée dans la section Préserver et archiver)

La confiance dans les résultats de la recherche repose, entre autres, sur le fait que les expériences ou les calculs soient reproductibles.

Au niveau technique, par exemple, la reproductibilité d'une mesure avec un même instrument, une méthode identique et dans un contexte donné, est essentielle pour valider les résultats d'une expérience. Au niveau scientifique, la répétabilité permet la validation des résultats obtenus. Il s'agit alors, par un autre moyen, d'arriver à des résultats équivalents.

Cependant, en fonction des disciplines, il peut être très compliqué de reproduire ou répliquer des résultats. La section suivante est destinée à permettre de mieux appréhender les enjeux, les défis et aussi les différentes facettes de la reproductibilité et de la répétabilité en fonction des disciplines et des méthodes mises en oeuvre.

11.1 Comprendre les enjeux et défis

Les présentations citées dans ce paragraphe concernent toutes divers aspects de la « reproductibilité » ou de la « répétabilité », qu'elles concernent les mesures, les accès aux données, les traitements, les calculs etc. . . . Chacune développe un point de vue particulier. Elles sont complémentaires.

Konrad Hinsén présente ici les enjeux et les défis de la recherche reproductible appliqués au calcul scientifique. Il reprend les bases de ce qu'est un calcul et de l'environnement dans lequel on utilise les programmes pour bien détailler les différents points de difficulté.

Enjeux et défis de la recherche reproductible

Vidéo : Konrad HINSEN, Centre de Biophysique Moléculaire Orléans et Synchrotron SOLEILAssemblée Aramis : La reproductibilité en pratique : méthodes et outils, 2019

David Hill fait ici la distinction entre reproductibilité et répétabilité et développe les difficultés de la répétabilité dans le calcul haute performance et en donne des exemples concrets dans différents domaines. Il propose des méthodes et des outils adaptés à chacun des points identifiés.

Reproductibilité et répétabilité - peut-on les négliger en calcul à haute performance ?

vidéo David Hill, Université Clermont Auvergne JCAD 2019

La présentation de Christophe Pouzat commence par un historique des raisons qui ont mené la communauté scientifique à prendre conscience, puis à prendre en compte le besoin de reproductibilité. La seconde partie de la présentation s'adresse plutôt aux développeurs qui produisent des codes.

Une brève histoire de la recherche reproductible et de ses outils

Christophe Pouzat, MAP5, Univ. Paris-Descartes et CNRS UMR 8145 CANUM 2016 : mini-symposium « Recherche reproductible », 2016

L'exemple présenté par Thomas Denecker concerne la biologie computationnelle. Il a pour but de sélectionner des gènes qui ne se comportent pas de la même façon entre deux conditions expérimentales. Les fonctionnalités présentées ne sont pas dépendantes de cet exemple. En effet, elles peuvent être appliquées à n'importe quelle autre question biologique. Brièvement, il s'agit de récupérer les données depuis des bases de données publiques, de réaliser une analyse reproductible avec un système de workflow dans un environnement virtuel dont l'ensemble du code, versionné, est disponible en open source. La visualisation des résultats est dynamique et un rapport en pdf ou html est disponible. Il regroupe les résultats de l'analyse et détaille l'ensemble des paramètres choisis par l'utilisateur.

La reproductibilité au service de la Biologie Computationnelle

Vidéo : Thomas Denecker Assemblée Aramis : La reproductibilité en pratique : méthodes et outils, 2019

11.2 Utiliser des environnements et des outils qui favorisent la reproductibilité

Parvenir à la reproductibilité peut être facilité par l'utilisation d'environnements et d'outils conçus ou adaptés dans cet objectif. De nombreuses équipes travaillent en ce sens depuis plusieurs années et plusieurs exemples sont présentés ici.

VIP, the « Virtual Imaging Platform », est un portail qui permet à ses utilisateurs d'accéder simplement à leurs données, de les traiter facilement avec des logiciels pré-installés sur la plateforme. Traitements et données sont distribués sur l'infrastructure EGL.

Virtual Imaging Platform : pour une science ouverte et reproductible

Vidéo : Sorina Camarasu-Pop, Frédéric Cervenansky, CREATIS JoSy 2019

Ludovic Courtès explique comment utiliser Guix et Jupyter pour la science reproductible. Guix est utilisé pour rendre l'environnement logiciel du notebook reproductible et déployé de façon automatique et reproductible. Nous vous conseillons de visionner la vidéo très didactique !

Vers un environnement reproductible pour les blocs-notes Jupyter

vidéo Ludovic Courtès, Inria JCAD 2019

Etre capable de reproduire des campagnes de calcul nécessite de connaître et savoir utiliser des outils adéquats. La bibliothèque [Python Execo](#) et le logiciel [OpenMole](#) sont deux exemples permettant de réaliser des campagnes de calcul reproductibles pour des applications de modélisation, simulations paramétriques, benchmarking, analyses de données numériques ou expérimentales.

Execo

Mathieu Imbert, Laurent Pouilloux Journées Campagnes de calcul reproductibles, 2018

Openmole (format tar)

Romain Reuillon, Mathieu Leclaire Journées Campagnes de calcul reproductibles, 2018

11.3 Développement open source et reproductibilité

Ces dernières années, les journées, ateliers traitant de la reproductibilité et du développement logiciel ont été nombreuses. Il est difficile d'isoler quelques présentations de ces ensembles pensés pour traiter d'une thématique d'ensemble. Aussi, dans cette section, plusieurs événements organisés par différents réseaux sont présentés globalement.

Lorsqu'on développe un code de calcul, seul ou à plusieurs, il est primordial de vérifier que chaque modification ne produit pas de régression dans l'ensemble de l'application. Il est donc nécessaire d'employer des tests unitaires, des tests d'intégration ou des tests du système complet. Ces tests s'intègrent dans un système de gestion de versions pour sauvegarder les modifications.

L'intégration continue fournit des outils permettant de valider l'intégrité du code à chaque soumission de modifications via github, gitlab, etc. Si, auparavant, il était assez fastidieux de mettre en oeuvre et d'administrer une chaîne d'intégration continue, les outils actuels sont très faciles à déployer. Ils offrent de plus des fonctionnalités qui permettent d'aller bien plus loin que la simple exécution de tests : la couverture du code, la validation syntaxique, la construction d'images de conteneurs et leur déploiement sur un dépôt, etc.

En 2017, le réseau « calcul » a organisé un « [Atelier intégration continue](#) » visant à se familiariser à trois outils d'intégration continue : *Jenkins*, *Travis CI* et *Gitlab CI*, en commençant par une introduction à deux outils couramment utilisés dans une chaîne d'intégration continue : « Git » pour gérer les versions et les publier vers un dépôt distant, puis « Docker » pour exécuter les tâches de compilation et de tests.

Les utilisateurs du langage de programmation « Julia » sont bien sûr également concernés par la reproductibilité. Ils trouveront dans la présentation de Mathieu Besançon à l'atelier « [JuliaNantes](#) », les réponses aux questions qu'ils peuvent se poser sur les raisons de veiller à la reproductibilité de leurs travaux et comment utiliser les outils Julia pour la science reproductible.

Getting started with Julia tools for reproducible science

Mathieu Besançon, Centrale Lille et Polytechnique Montréal Atelier JuliaNantes, 2019

Lorsque la reproductibilité n'est pas garantie, la validation, la vérification des logiciels, le processus de développement doivent être abordés différemment. On doit être en mesure d'estimer la précision des résultats numériques d'un logiciel et mettre en place des solutions pour contenir les sources d'imprécision. Lors de l'école [Précisions, Reproductibilité en Calcul et Informatique Scientifique](#) quatre thèmes en lien avec ces problématiques ont été abordés :

- Arithmétique flottante
- Mesurer la précision
- Recherche reproductible et calcul numérique

- Le calcul parallèle et le HPC sont-ils compatibles avec les enjeux de la recherche reproductible ? Pour chaque thème, des cours, présentations et TPs ont été organisés. Ces contenus répondront aux besoins des professionnels du calcul scientifique.

Les usines logicielles et les outils de production de code comme supports des bonnes pratiques de génie logiciel est une [thématique importante des JDEV 2020](#). Au cours de son exposé, Arnaud Legrand présente les différents enjeux, les solutions émergentes, les outils et standards, les plateformes qui permettent de tracer les codes, les simulations, les données...

Software Factories for Reproducible Big Data/AI/...

Arnaud Legrand, InriaJDEV 2020

Pour aller plus loin il peut être utile de consulter l'ouvrage collectif « Vers une recherche reproductible : Faire évoluer ses pratiques » cité dans la section [Autres guides de bonnes pratiques](#) de ce guide.

Autres guides de bonnes pratiques

Pierre-Yves Arnould et Marie-Christine Jacquemot. Guide de bonnes pratiques Gestion et valorisation des données de la recherche. [Rapport de recherche] OTELo; INIST-CNRS. 2016. hal-01275841

Véronique Ginouvès et Isabelle Gras. La diffusion numérique des données en SHS - Guide des bonnes pratiques éthiques et juridiques. Presses universitaires de Provence, 2018, Digitales, 9791032001790. hal-01903040

Olivier Brand-Foissac, Laurette Chardon, Marie David, Maurice Libes, Gilles Requilé et Alain Rivet. Guide de bonnes pratiques pour les Administrateurs Systèmes et Réseaux.

Isabelle André-Poyaud, Sandrine Astor, Olivier Baude, Fabrice Boudjaaba, Gaëlle Bujan, Béatrice Collignon, Frédéric Dubois, Emmanuel Kessous, Lionel Maurel, Muriel Roger. InSHS - Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte, Guide pour la recherche. CNRS, juin 2019

Nicolas Becard, Céline Castets-Renard, Gauthier, Martin Dantant, Laurence Freyt-Caffin Nathalie Gandon, Caroline Martin, Andrea Martelletti, Alexandra Mendoza-Caminade, Nathalie Morcrette, Claire Neirac. Ouverture des données de recherche, Guide d'analyse du cadre juridique en France. décembre 2017

Loïc Desquilbet, Sabrina Granger, Boris Hejblum, Arnaud Legrand, Pascal Pernot, Nicolas Rougier; Facilitatrice : Elisa de Castro Guerra. Vers une recherche reproductible : Faire évoluer ses pratiques. 29 mai 2019. hal-02144142

13.1 Auteurs

Ce guide de bonnes pratiques sur la gestion des données dans les réseaux métiers, a été réalisé par :

- Christine Hadrossek : DDOR, réseau Renatis
- Joanna Janik : DDOR, réseau Renatis
- Maurice Libes : réseau SIST
- Violaine Louvet : réseau Calcul
- Marie-Claude Quidoiz : réseau rBDD
- Alain Rivet : réseau QeR
- Geneviève Romier : réseau rBDD

13.2 Relecteurs

- Pierre Brochard : réseau DevLog
- Dominique Desbois : réseau DevLog
- Emilie Lerigoleur : réseau SIST
- Caroline Martin
- Pierre Navaro : réseau Calcul
- **version 1.0 Janvier 2021**

