Journée d'étude normande sur les données de la recherche
3 décembre 2021

Les données de la recherche et les réseaux métiers

Marie-Claude Quidoz (CEFE/CNRS)

Plan de la présentation

- Qui suis-je ?
- Retour d'expérience :
  Axe transversal « Cycle de vie des données » au CEFE
- GT « Atelier Données »
- Apprentissage d'une langue commune
- Cartographie des actions des réseaux métiers autour la gestion des données
- Guide de bonnes pratiques sur la gestion des données de la recherche

Qui suis-je ?

Cursus et parcours professionnel

DUT Statistique

CNAM Ingénieure Informatique
base de données
modélisation d'application

Responsable technique de la
plateforme SIE

1981 1982 — 1998 1999 — 2008 2009

IEPE (UMR – INSHS) → UREC (UPS – DGDR) → CEFE (UMR – INEE)

Gestionnaire de données
Données provenant des douanes, de l'OCDE, d'Eurostat, ...

Ingénieure en sécurité des systèmes d'information

Architecte des systèmes d'information

Données collectées : données déjà existantes (corpus, archives, …)

Données produites, lors de campagnes de recherche (observations, mesures, …)

Plateforme Systèmes d'Information en Ecologie (SIE)

Données issues de l'observation

Système d'information géographique / Bases de données / Calcul scientifique

Conseil et assistance technique aux utilisateurs du CEFE / Formation et veille technologique / Diffusion et valorisation des données

Réseaux métiers

- Réseaux coordinateurs & correspondants sécurité

- Réseau Bases de Données (rBDD)
  - ✔ Membre du comité de pilotage (2014)
  - ✔ Membre du bureau -> Référente formation (2014)
  - ✔ Formatrice interne (conception de base de données)

- GT Atelier Données : Responsable fonctionnel
  - ✔ Calcul, DevLOG, DDOR, INIST, Medici, QeR, rBDD, Renatis, Relier, Resinfo, RIS, SIST

- Membres des réseaux métiers « Bap E »
  - ✔ Calcul, DevLOG, Resinfo, RIS

Centre d'Ecologie Fonctionnelle & Evolutive

Retour d'expérience : Axe transversal « Cycle de Vie des données »

Le CEFE, c'est ... (1/4)

- **Centre d'Écologie Fonctionnelle et Évolutive (CEFE)**
  - ✔ Institut écologie et environnement du CNRS (INEE)
  - ✔ Créé en 1961
- **Un grand laboratoire en écologie**
  - ✔ 294 personnes (dont 94 chercheurs ; 49 IT ; 76 doctorants/post doctorants)
- **Un projet à large spectre**
  - ✔ Comprendre la dynamique, le fonctionnement et l'évolution du vivant, de «la bactérie à l'éléphant» et «du génome à la planète»
- **Un grand producteur de données**
  - ✔ D'observation / expérimentales / computationnelles

Le CEFE , c'est ... (2/4)

- **4 départements scientifiques**
  - ✔ 16 équipes de recherche
- **6 plateformes techniques**
  - ✔ Documentation, Bibliothèque
  - ✔ Génomique, Écologie Moléculaire et Evolution Expérimentale (GEMEX)
  - ✔ Analyses Chimiques (PACE)
  - ✔ Programmes à Long Terme (PLT)
  - ✔ Terrain d'Expérience (TE)
  - ✔ Système d'Information en Écologie (SIE)
- **5 axes scientifiques transversaux** (création officielle en 2020)
  - ✔ Écologie et agronomie
  - ✔ Evolution expérimentale
  - ✔ Sciences et sociétés
  - ✔ Biodiversité numérique
  - ✔ Cycle de vie des données

Le CEFE , c'est … (3/4)

- **Établissement(s) de rattachement**
  - ✔ Tutelles : CNRS, Université Montpellier, IRD, EPHE
  - ✔ Partenaires : INRAE, SupAgro, Université Paul Valéry Montpellier 3
- **Institut(s) de rattachement**
  - ✔ INEE
  - ✔ Secondaire : INSHS
- **Contrat de recherche financés par des institutions publiques** (HCERES 01/01/2014 au 30/06/2019)
  - ✔ 23 contrats européens (dont 6 ERC) en tant que porteur
  - ✔ 111 contrats nationaux (dont 50% ANR) en tant que porteur
- **400 publications par an**

Le CEFE , c'est … (4/4)

- De nombreux modèles de plan de gestion de données
  - ✔ INRAE, IRD, UM, UMPV3
  - ✔ ANR, ERC
- De nombreux entrepôts de données
  - ✔ Institutionnel : INRAE, IRD, TGIR Huma-Num, INEE (en 2021)
  - ✔ Ministériel : Recherche Data Gouv (UM en 2022)
  - ✔ Thématique : GenBank
  - ✔ Multidisciplinaire : DRYAD
- De nombreux documents officiels
  - ✔ CNRS : feuille de route pour la science ouverte / plan données de la recherche
  - ✔ UMPV3, EPHE : charte science ouverte

Pourquoi créer un axe transversal ?

- Mieux valoriser les données de la recherche
- Faire monter en compétence le personnel
  - ✔ Face aux déluges de mots clefs / sigles / concept
- Fournir un point d'entrée unique
  - ✔ Des zones blanches / chevauchements entre plateformes
- Améliorer la qualité des données
  - ✔ Sur toutes les étapes du cycle de vie des données
- Amplifier la dynamique déjà initiée
  - ✔ «Libérez vos Publications !» 2018

Animatrices de l'axe transversal

- **Véronique Arnal** (biologiste moléculaire)
  - ✔ Équipe BEV / plateforme GEMEX
  - ✔ Plateforme ADN dégradé
  - ✔ Référente données expérimentales et terrain
  - ✔ DU «Gestion des données de la science – Scientific Data Management»

- **Anne Gorgeon** (documentaliste)
  - ✔ Plateforme Documentation, Bibliothèque
  - ✔ CIST, Réseau Doccitanist, Réseau Renatis
  - ✔ Référente publications et HAL

- **Marie-Claude Quidoz** (gestionnaire de données)
  - ✔ Plateforme Système d'Information en Écologie
  - ✔ Réseau rBDD, GT Atelier Données
  - ✔ Référente données d'observation et préservation des données

Des réalisations (1/4)

(pour répondre à des questions des producteurs de données)

La sauvegarde des données

Pour remplir cette rubrique des plans de gestion des données, vous aurez besoin d'informations qui dépendent de l'infrastructure mise en place au CEFE.

Si vos données sont sur un espace du serveur Maison

- La sauvegarde a lieu du lundi au vendredi à 23 heures.
- Les sauvegardes sont conservées 30 jours.
- La restauration des dossiers pour lesquels l'utilisateur possède des droits est possible par l'utilisateur lui-même et ce pendant 15 jours grâce à un simple clic droit sur le dossier à restaurer. La restauration des dossiers peut toujours être faite par le service informatique.
- La restauration peut être faite au niveau d'un dossier ou d'un fichier.

Si vos données sont sur votre poste individuel sur le domaine CEFE

- La sauvegarde des postes fixes Windows a lieu du lundi au vendredi à partir de 11 Heures sauf s'ils sont éteints. Pour les portables Windows & Macintosh, la sauvegarde a lieu dès qu'ils se connectent à internet et ensuite toutes les 4 heures. Pour les linux, la sauvegarde n'est pas encore en place.
- 21 sauvegardes sont conservées
- La restauration des dossiers est faite par le service informatique.
- La restauration peut être faite au niveau d'un dossier ou d'un fichier.

Des réalisations (2/4)

(pour se former dans un premier temps et avoir un exemple concret)

DMPs publics

PGD 1 : Suivi (fictif) de population de poissons dans le lac du Bourget

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : Marie-Claude QUIDOZ, VERONIQUE ARNAL, Anne Gorgeon

Affiliation du créateur principal : CNRS

Modèle du PGD : ANR - Modèle de PGD (français)

Dernière modification du PGD : 17/09/2020

Financeur : ANR

Numéro de subvention : ANR-2020-000-000

Résumé du projet :

Avertissement : il s'agit d'un PGD destiné à être utilisé à des fins pédagogiques. Le projet est fictif, toute ressemblance avec des projets existants ou ayant existé ne saurait être que fortuite

L'objectif principal est d'observer les conséquences sur la faune marine des actions de dépollution engagées depuis le milieu des années 1970 dans le lac du Bourget qui présentait un phénomène majeur d'eutrophisation.

Le lac du Bourget n'a pour l'instant fait l'objet que de rares campagnes de prélèvements non standardisés.

Des réalisations (3/4)

(pour essayer de s'y retrouver dans ce déluge de mots)

DEFINITIONS AUTOUR DU CYCLE DE VIE DES DONNEES
Sélection de plusieurs définitions sur des termes courants en Science Ouverte

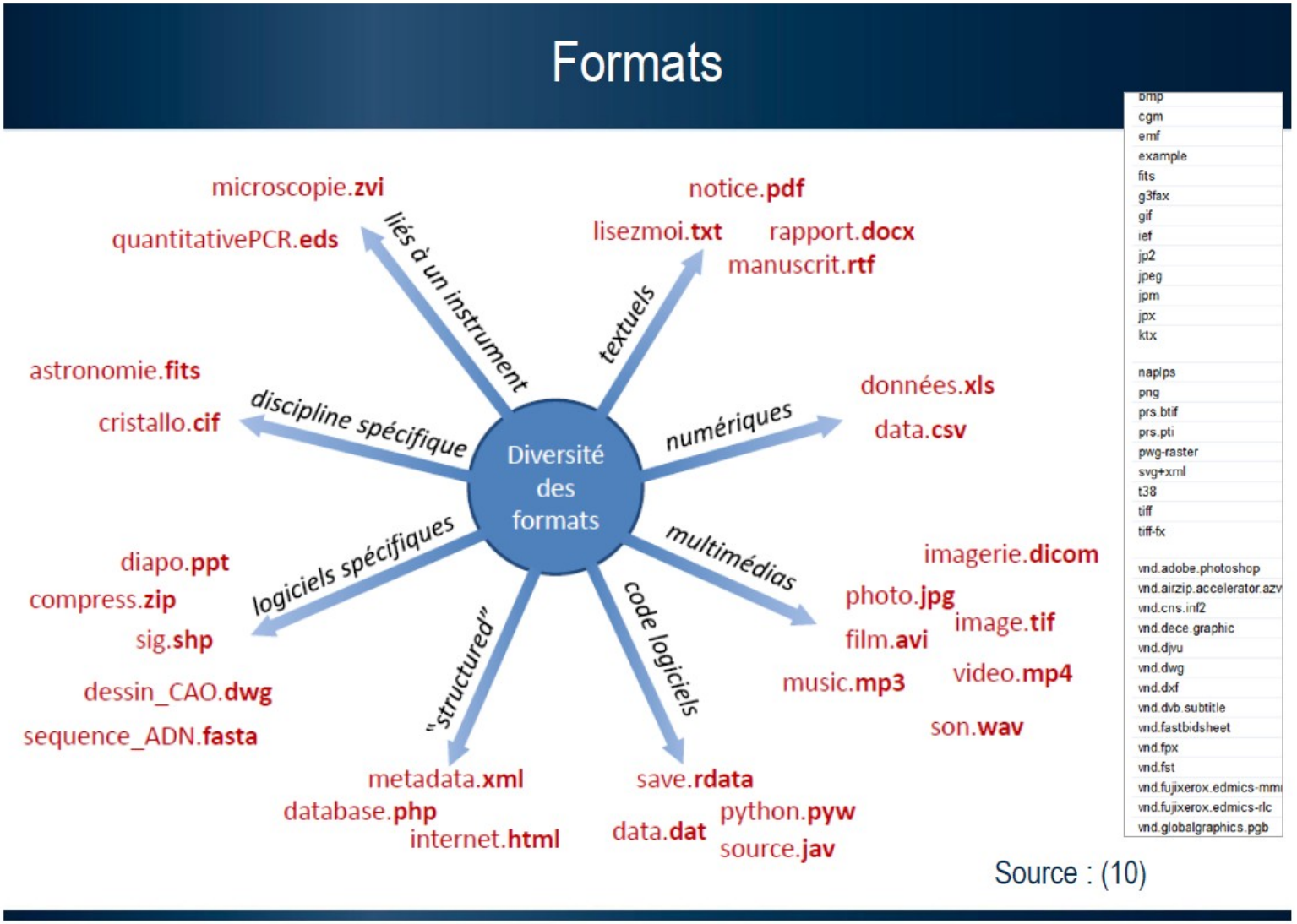
Des termes peuvent être ajoutés ou supprimés : merci de nous faire part de vos suggestions !

Axe : Cycle de vie des données
Auteur : Véronique Arnal
Relecteurs : Anne Gorgeon et Marie-Claude Quidoz
Mise à jour : 28/10/2020

Annuaire / Répertoire d'entrepôts (cf. Re3data, cf. OpenDOAR)
Site recensant des entrepôts de données, permettant de filtrer ses recherches par critère, comme re3data, OAD, OpenDOAR ; voir également l'annuaire d'entrepôts certifiés dans CoreTrustSeal.
Pour en savoir plus :
https://www.re3data.org/about
http://oad.simmons.edu/oadwiki/Data_repositories
http://v2.sherpa.ac.uk/opendoar/
https://www.coretrustseal.org/why-certification/certified-repositories/
Pour trouver un entrepôt, exemple de moteur de recherche Pangaea, Data Publisher for Earth & Environmental Science : https://www.pangaea.de/

Des réalisations en cours (4/4)

(pour apporter des conseils aux producteurs de données)

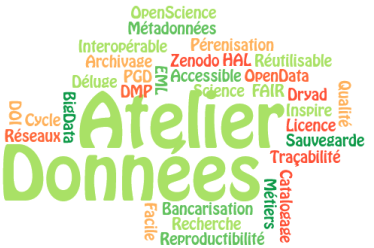


Méthode de travail

- **Réunions mensuelles**
  - ✔ Point sur les formations / séminaires suivis
  - ✔ Demandes internes
    - ► PGD, entrepôts de données, RGPD
- **Communication**
  - ✔ Intranet
  - ✔ Lettre Hebdomadaire
  - ✔ Intervention à des séminaires / formations
- **Objet d'étude fictif**
  - ✔ Suivi de population de poissons dans le lac du Bourget

Bilan au bout d'un an

- **Ne pas se démoraliser**
  - ✔ Le sujet est vaste
  - ✔ les compétences à acquérir sont nombreuses
  - ✔ Le sujet est en pleine structuration et évolue rapidement
  - ✔ Le personnel n'est pas fan
- **Ne pas se faire déborder**
  - ✔ On pourrait passer sa journée à suivre des webinaires
- **Se fixer une route et essayer de la suivre**
  - ✔ Développer l'acquisition de bonnes pratiques au CEFE
- **Respecter les passages obligatoires**
  - ✔ DMP Projet



GT « Atelier Données »

Genèse du GT « Atelier Données »

- **Rencontres des réseaux professionnels du CNRS (13-14/01/2016)**
  - Atelier : le jeu « inter-réseaux » : la donnée de l'acquisition à la valorisation»
  - 19 inscrits
  - Restitution : https://webcast.in2p3.fr/video/restitution_des_ateliers
  - De nombreux besoins recensés
- **Proposition de création du GT « Atelier Données »**
  - Mettre en place une journée thématique sur l'interopérabilité
  - Rédiger un guide pratique sur la traçabilité des activités de recherche
- **GT « Atelier Données»**
  - ✔ Une initiative des réseaux
  - ✔ Soutien de la Mission pour les Initiatives Transverses et Interdisciplinaires (MITI) du CNRS

Plate-forme réseaux de la « Mission pour les Initiatives Transverses et Interdisciplinaires » (MITI)

Les différents réseaux métiers et technologiques de la MITI ont pour point commun de fédérer une population autour d'un métier ou d'une technologie.

Les réseaux et la plateforme en quelques chiffres :

- 20 réseaux nationaux et près de 13000 adhérents (dont ~30% de chercheurs et chercheuses) et 59 réseaux régionaux associés,

De par leurs missions, ils répondent aux besoins des communautés scientifiques :

- participent à la réflexion et à la mise à disposition des outils, méthodes et infrastructures en matière de gestion et de partage des données scientifiques,
- conseillent et mettent en place de bonnes pratiques,
- organisent des formations et journées d'études. ~50 journées thématiques nationales ou ateliers techniques par an, ~30 actions nationales de formation par an

Lettre de mission du GT « Atelier Données»

- Apporter une vision transversale de la gestion des données afin d'enrichir la pratique de chaque réseau dans le domaine des données et permettre le développement de la complémentarité entre réseaux

- Valoriser l'apport des expériences et expertises « métier » dans une vision transversale de gestion de données

- Sensibiliser les communautés professionnelles de l'appui à la gestion des données (organisation de journées thématiques par exemple) ;

- Identifier les problématiques concernant les « data » dans chaque réseau (livrables à définir).

- Mettre en commun et partager de nouvelles pratiques en réseau et au sein de chaque réseau.

GT « Atelier Données » aujourd'hui

- 20 membres
- Deux duo d'animateurs
  - ✔ Pierre Brochard & Marie-Claude Quidoz (Juillet 2019 – Juin 2022)
- Différentes « structures »
  - ✔ Réseaux MITI : Calcul, Devlog , Medici, QeR, rBDD, Renatis, Resinfo, RIS
  - ✔ Autres réseaux : Relier, SIST
  - ✔ Structures : DDOR, INIST
- Des moyens de communication
  - ✔ https://gt-atelier-donnees.miti.cnrs.fr/
  - ✔ donnees-inter-reseaux@services.cnrs.fr (création en décembre 2019)
    - ► 400 adhérents – 350 messages
  - ✔ Lettre d'information mensuelle (la première a été faite en juin 2020)

Développement des sigles

- Calcul : réseau pour lacommunauté du calcul
- Devlog : résau national des développeurs en logiciel
- DDOR : direction des données ouvertes de la recherche
- INIST : institut de l'information scientifique et technique
- Medici : réseau des métiers de l'édition
- QeR : réseau qualité en recherche
- rBDD : réseau bases de données
- Relier : réseau qualité en ESR
- Renatis : réseau des professionnels de l'information scientifique
- Resinfo : réseau des administrateurs systèmes et réseaux
- RIS : réseau interdisciplinaire en statistique
- SIST : réseau des gestionnaires de données environnementales

Actions réalisées

(séminaire, webinaire, hackathon, FAQ, guide, cartographie)

- Cartographie des actions des réseaux métiers autour de la gestion des données (2017)
- Interopérabilité et pérennisation des données de la recherche. Comment FAIR en pratique ? Retours expériences (2018)
- Comment améliorer le dépôt de données de recherche (2020)
- Data paper, une incitation à la qualification et à la réutilisation des jeux de données (2020)
- Guide de bonnes pratiques sur la gestion des données de la recherche (2021)
- Qualité des données (2021)
- Foire Aux Questions (entrepôts de données et les data papers)

Apprentissage d'une langue commune

Des mots aux multiples définitions

(données de la recherche)

- « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche » (OCDE, 2007).

- « ensemble des informations scientifiques produites ou collectées dans le cadre d'un projet de recherche, les données de la recherche peuvent être des photos, des mesures, des sons, etc. Elles sont nécessaires comme éléments probants afin de valider les résultats de la recherche et doivent être accompagnées d'informations qui les documentent, telles que des protocoles expérimentaux, des descriptifs méthodologiques ou des métadonnées. Ces données peuvent être diffusées dans des entrepôts généralistes ou spécialisés » (Consortium COUPERIN).

- Plan données de la recherche CNRS (2020) : « les données de la recherche sont les données brutes ou retraitées dans tous les formats, les textes et documents et également les logiciels, les algorithmes, les protocoles et les workflows

Des mots compris différemment (1/3)

(stocker, sauvegarder, archiver, préserver, pérenniser)

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

https://mi-gt-donnees.pages.math.unistra.fr/guide/06-archiver.html

Des mots compris différemment (2/3)

(stocker, sauvegarder, archiver, préserver, pérenniser)

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-CNRS
Séminaire « 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, ISTerre
MITI-CNRS, 2018

https://mi-gt-donnees.pages.math.unistra.fr/guide/06-archiver.html

Des mots compris différemment (3/3)

(stocker, sauvegarder, archiver, préserver, pérenniser)

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 chapître, nous utiliserons les termes "préserver / préservation" qui sont les termes le plus utilisés actuellement.

https://mi-gt-donnees.pages.math.unistra.fr/guide/06-archiver.html

Des mots compris différemment
(métadonnées)

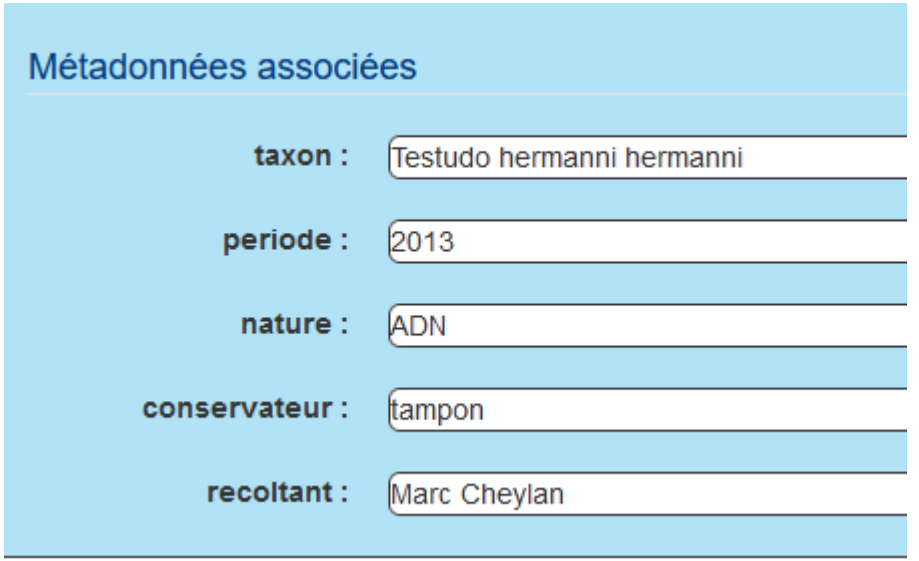
Les métadonnées sont des données décrivant d'autres données. Un enjeu majeur de l'Open Data pour la réutilisation des données de la recherche est la description standardisée de celles-ci.

https://doranum.fr/metadonnees-standards-formats/

EXEMPLE

- **Dublin Core** *(interdisciplinaire), description des ressources numériques.*
- **MARC** *(Machine-readable cataloging), description du contenu des bibliothèques.*
- **EAD** *(Encoded Archival Description), description des archives.*
- **DwC** (*Darwin Core*), *domaine de la biodiversité.*
- **DDI** (*Data Documentation Initiative*), *domaine des sciences sociales, comportementales et économiques.*
- **EXIF (***Exchangeable image file format*), *description technique et automatique d'un cliché.*
- **IPTC** (*International Press Telecommunications Council*), *description d'une image par l'auteur.*

https://doranum.fr/metadonnees-standards-formats/fiche-synthetique/



https://collection.cefe.cnrs.fr/

Des mots compris différemment

(entrepôt de données)

Service en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données.

≠

Entrepôt de données ou EDD (ou **base de données décisionnelle** ; en anglais, **data warehouse** ou DWH) désigne une base de données utilisée pour collecter, ordonner, journaliser et stocker des informations provenant de base de données opérationnelles et fournir ainsi **un socle à l'aide à la décision** en entreprise.
Wikipédia

https://e-envir.sciencesconf.org/data/pages/J4_CM8_Entrepot_de_donnees_Jean_ChristopheDesconnets.pdf

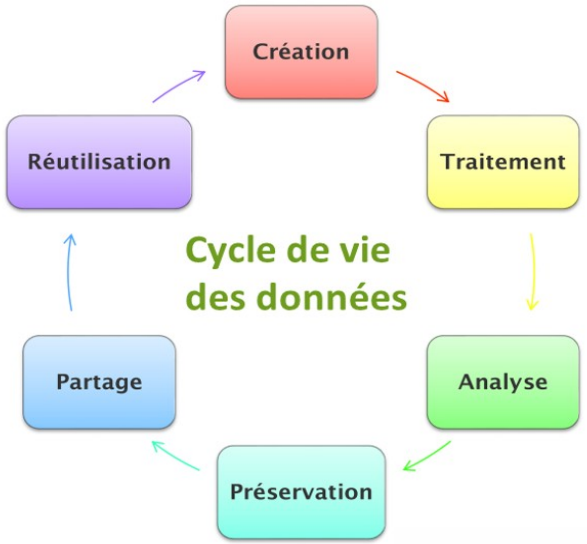
Des buts compris différemment

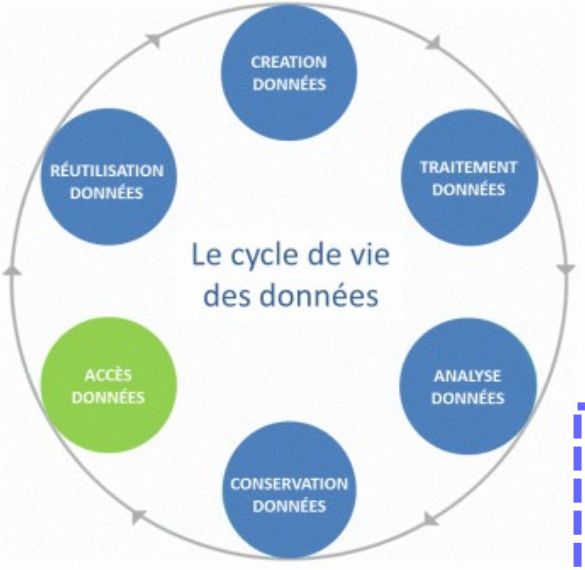
(entrepôt de données)

- Partage ou archivage (pérenne) ?
- La confusion provient peut-être
  - ✔ De la notion d'identifiant pérenne qui est associé aux jeux de données lors du dépôt dans un entrepôt de données
  - ✔ Du terme « partage » qui est mal compris
  - ✔ Du terme « pérenne » qui ne semble pas avoir la même définition pour tout le monde
    - ► 5 ans ? 10 ans ? 50 ans ?
- Etape « publier » et/ou « préserver » d'un cycle de vie des données ?

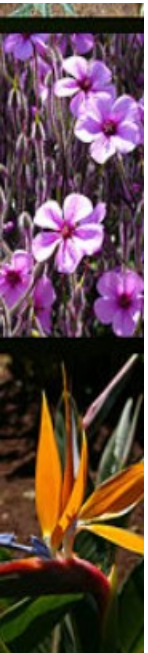
Cartographie des actions des réseaux métiers autour la gestion des données

Définir notre cycle de vie des données

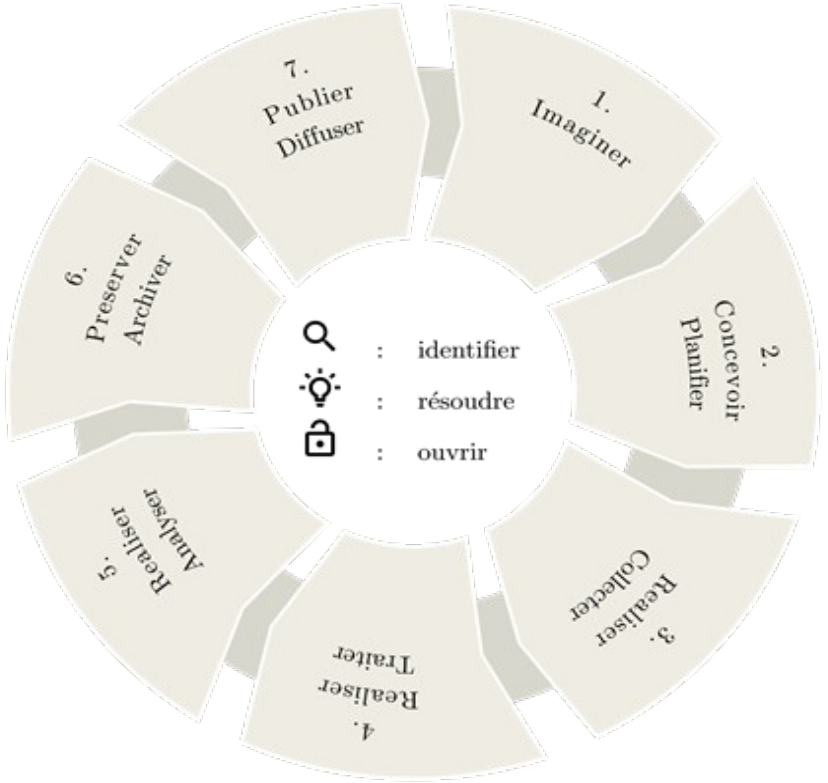




GT « Atelier Données »



Se définir sur chaque étape



- **Identifier : interrogations principales concernant les données**
- **Résoudre : actions et solutions mises en oeuvre**
- **Ouvrir : manques ou perspectives à explorer**


3. Réaliser et collecter
CALCUL
Assurer les développements collaboratifs, contribuer aux bases de données communautaires
Collecter les données brutes, obtenir des métadonnées riches, utiliser des normes et des standards ouverts
Articuler méthodologies des infrastructures, standards communautaires et besoins de référentiels sémantiques pour les métadonnées
DEVLOG
Intégration de données externes (produites par des utilisateurs et/ou des logiciels connexes)
Définir des dispositifs de collecte adaptés
Pouvoir connecter des systèmes de données externes
QeR
Maitriser et contrôler l'acquisition des données
Mise en œuvre de bonnes pratiques : étalonnage et suivi des équipements, validation des logiciels, renseignement du cahier de laboratoire, identification et conservation des échantillons
Identification systématique des échantillons
MEDICI
Choix de standards de métadonnées et mise en conformité des métadonnées existantes avec les exigences liées aux supports de publication et/ou aux contraintes des entrepôts
Participer au travail d'appropriation et d'adaptation des langages informatiques existants, former les éditeurs
Veille sur les formats et accompagnement de leur appropriation
RBDD
Variété des modalités d'acquisitions de données (observation, enquêtes, capteurs, interfaces)
Participer à l'élaboration des méthodologies de collecte de l'information, conseiller sur le choix des référentiels pour les métadonnées
Développement de procédures d'intégration de données externes dans les bases de données
RENATIS
Constituer des jeux de données pertinents en prenant en compte leur potentiel d'impact dans les communautés scientifiques et pour le dialogue science / société
Mettre en œuvre les plans de gestion de données en s'appuyant sur une identification de l'existant, l'organisation des fichiers et le versionnage
Utiliser des référentiels qualité cohérents avec les objectifs stratégiques des projets
RESINFO
Résoudre les questions liées à la gestion des données : volumétrie, infrastructure de stockage, transfert de données, débits, interopérabilité
Description des processus de production et de gestion des données dans des plans de gestion de données et mise en place de chaînes d'acquisition et de transfert "du capteur au serveur"
Participer à la définition et la rédaction des plans de gestion de données avec les chercheurs

Intérêt de ce travail

- S'approprier les étapes du cycle de vie des données
- Mieux connaître le coeur de métier de chaque réseau
- Identifier les manques et les zones de recoupement entre les réseaux
- Initier des collaborations entre réseaux
  - ✔ Interventions dans des formations

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

Originalité

- Son originalité réside dans son application aux données de la recherche sous l'angle des pratiques de différents métiers de la recherche

- Il fournit un point de vue transversal à travers une compilation de diverses pratiques métiers. Il présente :
  - ✔ les nombreuses actions de formation ou de sensibilisation des réseaux ;
  - ✔ les compétences et expertises développées issues de pratiques standardisées qui font leurs preuves sur le terrain ;
  - ✔ des recommandations et des solutions techniques et organisationnelles grâce à la veille technologique et juridique réalisée très régulièrement.

- Il traduit les efforts et le soutien mis en place par les membres des réseaux, dans la gestion et la valorisation des données scientifiques.

Quelques extraits (1/3)

Guide de bonnes pratiques • Atelier Données •

Rechercher dans ce livre ...

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, CERMAV*
> *Atelier 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.

Quelques extraits (2/3)

Guide de bonnes pratiques – Atelier Données

Rechercher dans ce livre ...

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, LLL
Journée « 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.

Quelques extraits (3/3)

Guide de bonnes pratiques • Atelier Données •

Rechercher dans ce livre ...

7.4.3. Retour d'expériences d'utilisation de DOI

Philippe Techiné 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 Téchiné, 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 Téchiné, 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".

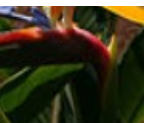
Retour d'expérience sur l'attribution de DOI à l'OSU OREME.

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 Quidoz 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 Quidoz, atelier traçabilité RBDD 2018



Pour en arriver là

- Des productions dans les réseaux métiers
  - ✔ Guide de bonnes pratiques pour les administrateurs systèmes et réseaux (Resinfo 2012)
- Des initiatives en gestation
  - ✔ Guide de bonnes pratiques dans le domaine des bases de données (rBDD)
  - ✔ Guide de bonnes pratiques la gestion des données d'observation (SIST)
- En juillet 2019, constitution d'un groupe de personnes motivées pour se lancer dans l'aventure
- Un écueil : ne pas faire le n^{ième} guide sur ce thème
- Une évidence : s'appuyer sur le cycle de vie des données

Contenu : cycle de vie des données

Pour adopter un point de vue commun aux différents métiers et activités de nos réseaux :

- **Nous nous basons sur le cycle de vie des données**
- **le cycle de vie des données représente un cadre structurant et fournit un vocabulaire commun**

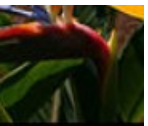
Le guide fournit une lecture nouvelle des actions des réseaux, enrichie des approches complémentaires des pratiques des différents réseaux

1 . Imaginer - Préparer

"Imaginer" est la première étape de notre cycle de vie des données.

- phase *préparatoire* qui correspond à *l'identification des problématiques techniques et juridiques* associées à la gestion des données

- L'apport des réseaux est ici important en termes de croisement des disciplines et des métiers pour *apporter un éclairage global et répondre au mieux aux besoins des communautés scientifiques :*
  - s'informer, comprendre pour anticiper et envisager le déroulement d'un projet.
  - connaître les *contraintes et opportunités, les outils et infrastructures disponibles, les politiques d'accompagnement, les acteurs, les réglementations en vigueur ou encore les compétences et expertises à acquérir.*

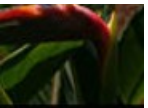


2. Concevoir - Planifier

Dans cette étape, on *définit les tâches à accomplir pour réaliser le projet de recherche, élaborer un planning, rechercher d'éventuels partenaires et financements, et élaborer les spécifications nécessaires*

Pour ces travaux de conception et de planification, les réseaux *apportent un appui sur la gestion et les méthodologies de conduite de projet*, et conseillent et mettent en place *des outils pour assurer* ***l'interopérabilité*** *des systèmes mis en oeuvre :*

- Recommandations et des retours d'expérience pour *commencer la rédaction de plans de gestion de données (DMP)*
- I*dentification des infrastructures adaptées au projet* (fonctionnalités, capacités et services fournisseur du service)
- *Mise en place du mode de collecte et de stockage* afin d'organiser la traçabilité en amont, traçabilité qui permettra de garantir la réutilisation des données



3. Collecter

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, avec leurs métadonnées descriptives.

Il s'agit donc, dans cette phase :

- de *travailler sur les processus d'acquisition des données* obtenues : capteurs environnementaux, instruments, sondages, modèles numériques
- d'assurer la traçabilité des données : cahiers de laboratoires, tablettes de terrain…
- de rendre ces données « FAIR » en les décrivant et en y associant des métadonnées, en *utilisant des normes et des standards (thésaurus, vocabulaire contrôlés…) afin que les données soient interopérables*
- se prémunir des pertes, en stockant et sauvegardant les données

4. Traiter

Cette phase correspond au *prétraitement des données brutes issues des acquisitions et des collectes.*

Il s'agit souvent de :

- *regrouper, choisir, qualifier les données pertinentes* puis les *transformer dans des formats standards interopérables*, et les préparer en vue de leur analyse ultérieure.
- Utiliser des infrastructures logicielles , services d'intégration de données ("*framework*"), 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

5. Analyser

L'étape d'analyse des données correspond à *l'extraction de l'information des données traitées.*

Cela recouvre de nombreux types de techniques : *calcul intensif, traitement statistique, machine learning, visualisation* …, ce qui peut nécessiter également des plateformes de traitement adaptées.

Cette étape du cycle de vie *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.

6. Préserver - Archiver

Sauvegarder, préserver, sécuriser l'information et, voire archiver les données sont des phases essentielles de la gestion rigoureuse des données.

Les notions de *stockage, de sauvegarde et d'archivage* ainsi que les actions de *préservation et de pérennisation* revêtent des notions et des sens et des pratiques différentes que nous explicitons dans le Guide.

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

7. Publier et Diffuser

Cette étape consiste à *publier et diffuser les données de manière à ce qu'elles soient accessibles et réutilisables selon des formats et des processus interopérables.*

L'accompagnement des réseaux s'exerce sur :

- l*e processus de publication des données dans des "**catalogues**", des "**entrepôts"** ou des plateformes techniques, pour en permettre l'accès*,
- la documentation des données avec des métadonnées descriptives provenant de vocabulaires contrôlés et de leurs formats d'exploitation pour en assurer la réutilisabilité.
- 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
- l'*identification des données via des **identifiants pérennes**, lors du dépôt dans des entrepôts de données.*
- la publication de "*Datapaper*" pour valoriser et expliciter en détail les données

Un guide pour qui ?

- Ce guide n'est pas exhaustif puisqu'il est le reflet des thèmes abordés dans le cadre des actions des réseaux impliqués dans la rédaction du guide.
- Il est peut être trop centrée donnée et pas assez logiciel bien qu'il y ait un chapitre sur la reproductibilité
- Un guide pour tous afin de mieux connaître l'apport des métiers d'appui à la recherche dans la gestion des données de la recherche
- Une « mine d'or » pour les métiers d'appui à la recherche pour s'informer et se former

Crédits (version 1.0 Janvier 2021)

Production d'un sous-groupe du GT « Atelier Données »

Auteurs

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

Relecteurs

- Pierre Brochard : réseau DevLog
- Dominique Desbois : réseau DevLog
- Emilie Lerigoleur : réseau SIST
- Caroline Martin : réseau RELIER
- Pierre Navaro : réseau Calcul

Edition Web

- Pierre Navaro : réseau Calcul

Conclusion

Que de chemin parcouru !

Janvier 2016 -> Décembre 2021

- **En évolution permanente**
  - ✔ DMP -> MaDMP
  - ✔ Entrepôt de données -> Entrepôt de données certifié
  - ✔ Veille technologique importante
- **Au périmètre qui s'étend**
  - ✔ Logiciel
  - ✔ Langue commune à acquérir
- **Qui demande à avoir un pied sur le terrain**
  - ✔ Ne pas proposer des solutions inaplicables

Merci de votre attention

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