Groupe de travail inter-réseaux sur les données

soutenu par la Mission pour l'Interdisciplinarité (MI) du CNRS

Anne Cadiou

Réseau Calcul

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

mercredi 5 juillet 2017

Vandoeuvre-lès-Nancy du Globe de Paris

Mission interdisciplinarité
Centre national de la recherche scientifique

Historique

- **Réunion le 15 avril 2015** à l'initiative de Renatis
  - Identifier des thématiques communes autour des données
  - CORIST, DIST, Docplanet, Go!Doc, INSHS, MATE-SHS, Médici, Renatis, RNBM
- **Atelier Données les 13-14 janvier 2016**

  lors des journées de rencontre des réseaux professionnels organisées par la Mission pour l'Interdisciplinarité du CNRS
  - 19 participants : Calcul, Cogiter, Cristec, DevLog, Loops, Médici, rBDD, RCCM, Renatis, Resinfo, RTMFM, QeR...
  - **Création du groupe de travail inter-réseaux à la MI**

    piloté par

    Caroline Martin (Médici), Emmanuelle Morlock (Renatis)
- **Groupe de travail MI Données**
  - Calcul, DevLog, Médici, QeR, rBDD, Renatis, Resinfo
  - y ont participé : Romaric David, Loïc Gouarin, Anne Cadiou (Calcul), Lyriane Bonnet (DevLog), Caroline Martin, Stéphane Renault (Médici), Alain Rivet (QeR), Geneviève Romier, Marie-Claude Quidoz (rBDD), Colette Orange, Philippe Eyraud, Emmanuelle Morlock (Renatis), Olivier Brand-Foissac, Maurice Libes (Resinfo)

Enjeux identifiés

- **Répondre à l'évolution rapide** du travail sur les données dans le contexte actuel
  - Big Data (quantités massives, complexité) (défi Mastodons)
  - accessibilité, interopérabilité,
  - reproductibilité, réplicabilité, réutilisabilité,
  - préservation, pérennisation,
  - qualité...
- **Identifier les collaborations** possibles entre les réseaux

Objectifs du groupe de travail

- **Établir** et **diffuser** une vision transversale de la gestion des données afin d'enrichir la pratique de chaque réseau et permettre le développement de la complémentarité entre réseaux
- **Identifier** les problématiques concernant les données dans chaque réseau
- **Valoriser l'apport** des expériences et expertises **métier** entre les réseaux
- **Sensibiliser** les communautés professionnelles sur la gestion des données
- **Mise en commun et partage de nouvelles pratiques**

Méthode de travail

- Depuis janvier 2016 : 5 réunions en visio, 2 en présentiel
- Travail entre les sessions (préparation, relecture)
- Utilisation de pad, d'un wiki (gitlab, outils collaboratifs)
- Journée d'étude avec des experts externes
  (Francis André - DIST, Alain Bénard - INRA)
- Sondages internes aux réseaux
  sur leurs pratiques et leurs besoins

- Apprendre à **se comprendre**
  et à s'accorder sur le vocabulaire de chacun !
- **Réaliser une cartographie** du qui fait quoi sur les données
  pour chaque réseau participant

- Travailler sur le cycle de vie des données propres aux métiers
- Un **cycle unique**, des **données différentes**

Exemple avec le réseau Calcul

Définition de la donnée pour Calcul

désigne les **codes**, les **résultats produits ou analysés** par ces codes (par exemple issus de dispositifs expérimentaux), les **publications**, les **bases de données** (bases de référence, bibliothèques scientifique, benchmarks), etc.

Activités

- Créer/Acquérir les données
- Traiter les données
- Analyser les données
- Préserver les données
- Publier/Diffuser/Rendre accessibles les données
- Réutiliser les données

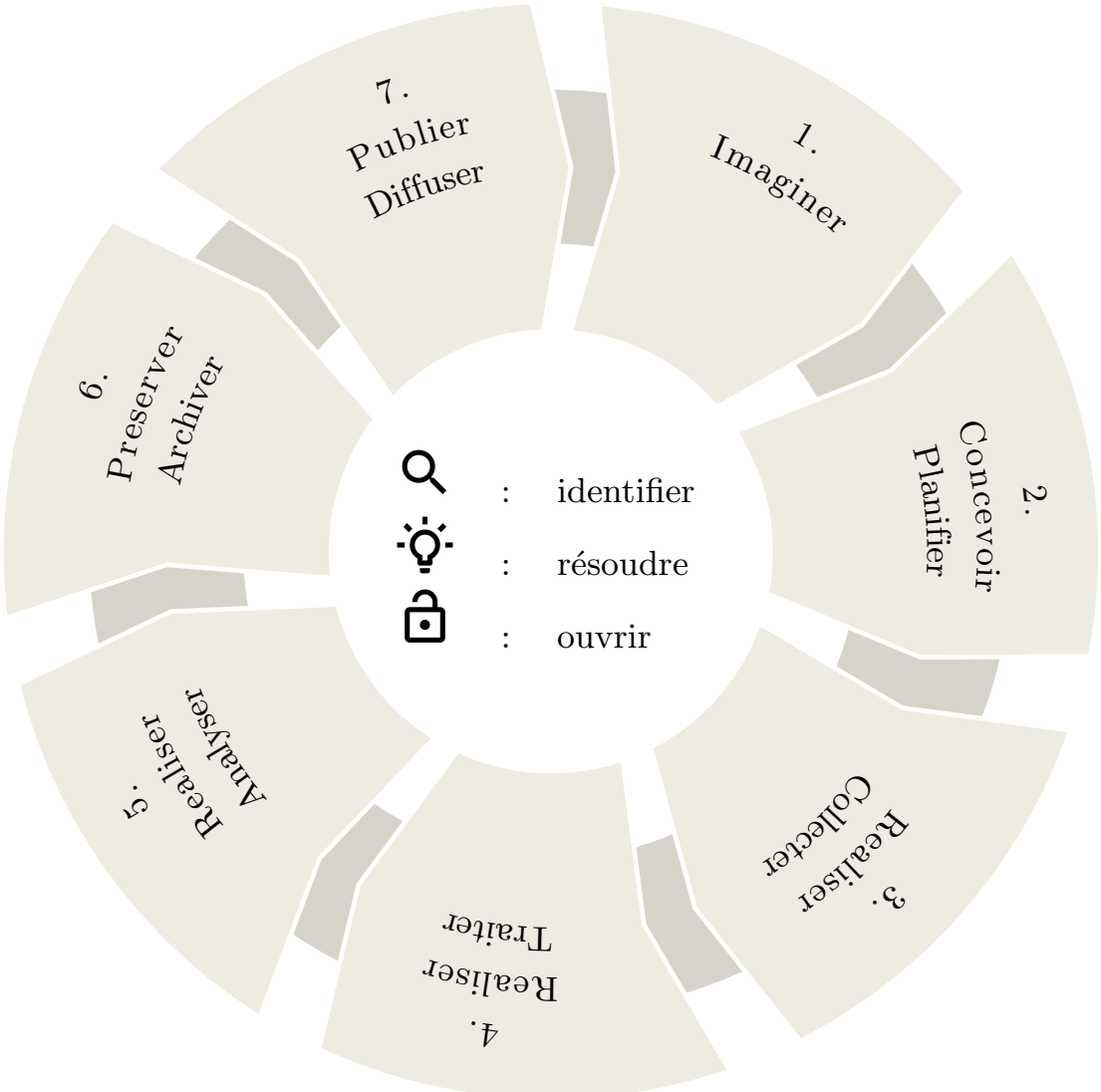
Extraits des résultats de l'enquête auprès du réseau Calcul

- **Type :** 50% binaire 50% ascii
  40% non dupliquées (10 Go - 100 To)
  79% à garder $\sim$ 5 ans
- **Méthode :** 78% n'identifient aucune structuration des données dans leur communauté

Enjeux identifiés

- stockage/sauvegarde pérenne sur le long terme des gros volumes de données (avec les codes et les processus d'analyse) - tracabilité, reproductibilité, réutilisabilité - définition de format de données lisibles sur le long terme,
- passage à l'échelle des outils de gestion et d'analyse des données,
- pouvoir faire des analyses in-situ sur les données,
- rendre possible l'interopérabilité des données et des processus d'analyse de celles-ci.

Cycle de vie



identifier : problématique de l'étape telle que perçue par le réseau

résoudre : solutions trouvées par le réseau sur cette étape

ouvrir : manques à combler ou défis et perspectives à explorer

1. Imaginer

Imaginer les données, c'est proposer de nouvelles solutions techniques ou technologiques pour répondre à une problématique de recherche scientifique.

L'**apport d'un réseau** consiste à accélérer le processus exploratoire en croisant les pratiques et favorisant la veille technologique sur les solutions en émergence.

Ce qui manque n'est pas encore complètement intégré dans les usages métier des membres d'un réseau. Ce manque peut être soit partiellement comblé en s'inspirant des solutions déjà en usage dans d'autres réseaux, soit à inventer pour trouver sa propre solution.

1. Imaginer - identifier, résoudre, ouvrir - par réseau

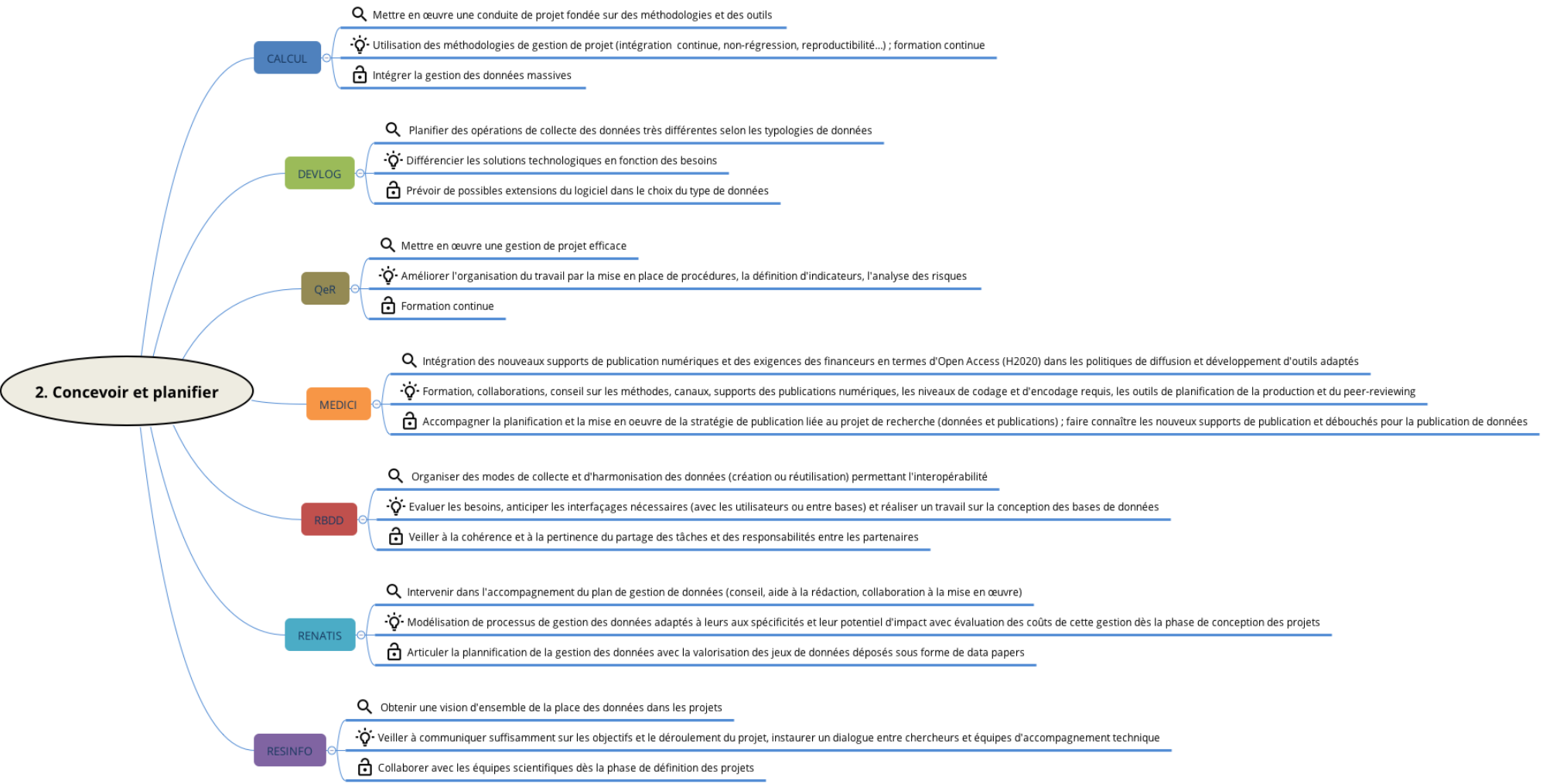
2. Concevoir et planifier

Chaque métier fonctionne en **mode projet** dans ses fonctions d'appui à la recherche. Les méthodes et solutions technologiques mises en oeuvre dépendent de l'analyse des besoins.

Un réseau métier apporte à chacun **une vision transversale** aux cas particuliers et enrichit les réponses aux besoins.

Pour fonctionner ainsi, chaque métier doit pouvoir interagir et **collaborer très en amont** dans les projets de recherche. Les réseaux peuvent contribuer à promouvoir cette façon de travailler.

2. Concevoir et planifier - identifier, résoudre, ouvrir -



3. Réaliser et collecter

Les métiers participent à la réalisation et la collection des données.

Les réseaux contribuent à **populariser les bonnes pratiques**, en intégrant les méthodes favorisant la traçabilité, la réutilisabilité, l'interopérabilité, autant que possible.

Sur ces aspects, une **vision inter-réseaux** pourrait **contribuer à faire émerger de nouveaux standards** méthodologiques et technologiques ou **à propager des solutions ad-hoc** pour différents usages.

3. Réaliser et collecter - identifier, résoudre, ouvrir -

4. Réaliser et traiter

Cela concerne le travail pour passer d'une donnée brute à une donnée raffinée (modélisation, reformatage, mise en base de données, métadonnées, documentation, contrôle qualité...)

Forte convergence des métiers **sur le principe** d'actions.

Possibilité de partage d'expériences, formations, entre-aide sur les outils émergeants (correction automatique...)

4. Réaliser et traiter - identifier, résoudre, ouvrir - par réseau

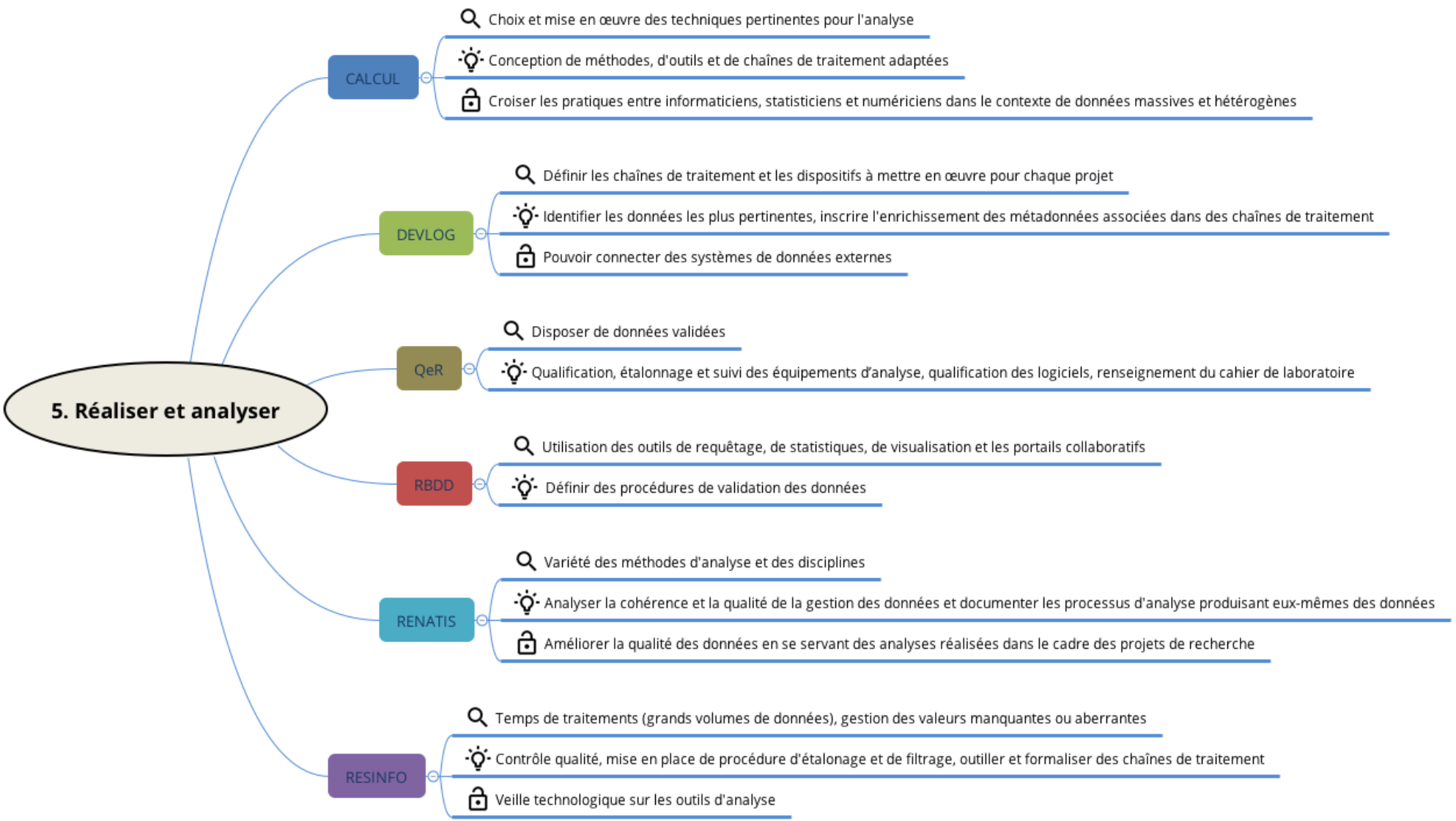
5. Réaliser et analyser

Les analyses réalisées diffèrent suivant les communautés et les métiers.

Les métiers convergent vers la **mise en place de chaînes de traitement**, développant ou exploitant les **outils et procédures adaptés**. Ils se préoccupent de contrôle qualité et de performance.

Encore une fois, le **partage de bonnes pratiques** entre réseaux métiers sur les méthodologies paraît possible, malgré les diversités applicatives.

5. Réaliser et analyser - identifier, résoudre, ouvrir -



6. Préserver et archiver

Tous les **métiers convergent** sur le fait d'**intégrer la pérennisation**, la traçabilité, l'accessibilité des données et des outils permettant de les exploiter.

Les réseaux métiers sont amenés à participer à la **définition des politiques de sauvegarde et d'archivage**, et à **popularisation des procédures** associées dans les collectifs de recherche.

Ces aspects sont **un dénominateur commun fort** de l'ensemble des réseaux. Des actions de retours d'expérience et sensibilisation à des méthodologies communes seraient un apport intéressant d'une vision inter-réseaux sur ces aspects.

6. Préserver et archiver - identifier, résoudre, ouvrir -

7. Publier et diffuser

Cette étape est indissociable de la mise à disposition de moyens, notamment en **infrastructures**. Cela s'accompagne aussi de supports à la diffusion (valorisation, aspects juridiques...)

Les différents métiers travaillent sur l'ensemble des informations (données, modes opératoires, échantillons, publications, visualisation et interfaces graphiques) nécessaires à la mise en oeuvre des supports de diffusion et de valorisation.

Des **savoir-faire complémentaires** existent dans les réseaux métiers sur ces différents aspects, qu'il pourrait être intéressant de partager.

7. Publier et diffuser - identifier, résoudre, ouvrir - par réseau

Actions récentes des réseaux autour des données

- rBDD, *Conduire et construire un plan de gestion des données*, ANF 2015
- rBDD/DevLog, *Bases de données NoSQL*, ANF 2015
- Calcul, *Données scientifiques massives : stockage et visualisation*, ANF 2016
- Resinfo, *Des données aux "Big Data" : exploitez le stockage distribué !*, ANF 2016
- Renatis, *Construire un projet de gestion de données de la recherche*, ANF 2016
- rBDD, *Comment concevoir une base de données*, ANF 2017
- Calcul, *Méthodes et outils du calcul pour la réduction de la dimension dans l'analyse de données massives*, ANF 2017
- Renatis-Médici, *Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données*, ANF 2017

Perspectives

- Partager les expériences métiers autour de la gestion des données (pérennisation...)
- Favoriser les collaborations inter-réseaux
  - Créer un réflexe d'interventions invitées dans les actions des réseaux (ex. Calcul en 2016)
  - Actions co-portées (ex. rBDD/DevLog en 2015)

Le réseau Calcul a identifié 3 sujets qui l'intéressent particulièrement :

- Techniques pour la science reproductible (cf école thématique 2013 'Précision et reproductibilité en calcul numérique')
- Conseils pour la rédaction d'un Data Management Plan (à l'échelle d'un laboratoire, d'une plate-forme)
- Méthodes et outils pour l'analyse des données de très grande taille