



## LA SCIENCE OUVERTE : vers les connaissances partagées

Comment rendre les données FAIR en s'appuyant sur le guide de bonnes pratiques sur la gestion des données de la recherche de l'Atelier Données ?

19 – 20 octobre 2021, Meudon

Alain Rivet (CERMAV)

Geneviève Romier (CC IN2P3)



# Plan de l'atelier

Les données de la recherche

L'atelier-données (MITI)

Génèse du guide

Les chapitres du guide

Interactivité



# Pyramide des données

La pyramide de publications des données



Tutoriel Form@doc (Schéma adapté de Report on integration of data and publications. Opportunities for Data Exchange (Reilly S. et al., 2011))

LA SCIENCE OUVERTE : vers les connaissances partagées



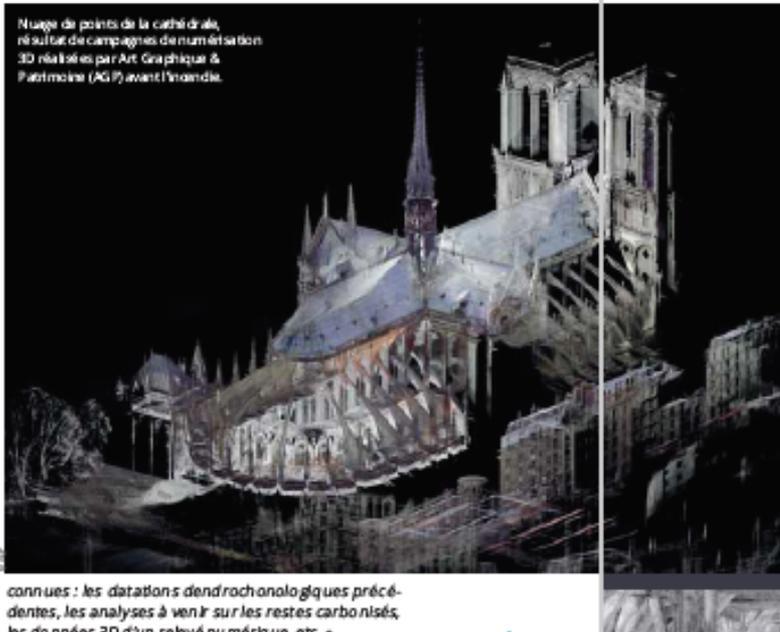
# Notre dame

## Notre-Dame et son double numérique

Créer une sorte de Google Earth de la cathédrale de Paris, tel est l'ambitieux projet d'une équipe de chercheurs. Objectif : regrouper au sein d'une plateforme collaborative la totalité des connaissances passées et à venir sur l'édifice. Explications.

« La restauration de Notre-Dame de Paris sera un chantier historique. » Et Livio de Luca, directeur du laboratoire Modèles et simulations pour l'architecture et le patrimoine (MAP), et lauréat 2019 de la médaille de l'innovation (lire p. 10), ne souhaite pas que cette histoire se perde à nouveau dans les flammes. « Nous allons créer un système d'information intégrant toutes les données scientifiques et techniques sur la cathédrale », explique-t-il.

Le groupe de travail qu'il coordonne s'apprête ainsi à en réaliser une sorte de « double numérique » rassemblant tout ce que l'on sait de l'édifice, des croquis de construction jusqu'au relevé 3D de son état actuel, et qui sera capable d'intégrer toute l'information à venir. En fait d'une simple réplique en images de synthèse, il s'agit plutôt de construire une base de données et de connaissances inédite. « Grâce à elle poursuit-il, les différentes équipes du chantier pourront partager leur expertise et prendre in fine de meilleures décisions. »



Nuage de points de la cathédrale, résultat de campagnes de numérisation 3D réalisées par Art Graphique & Patrimoine (AGP) avec l'Inaodis.

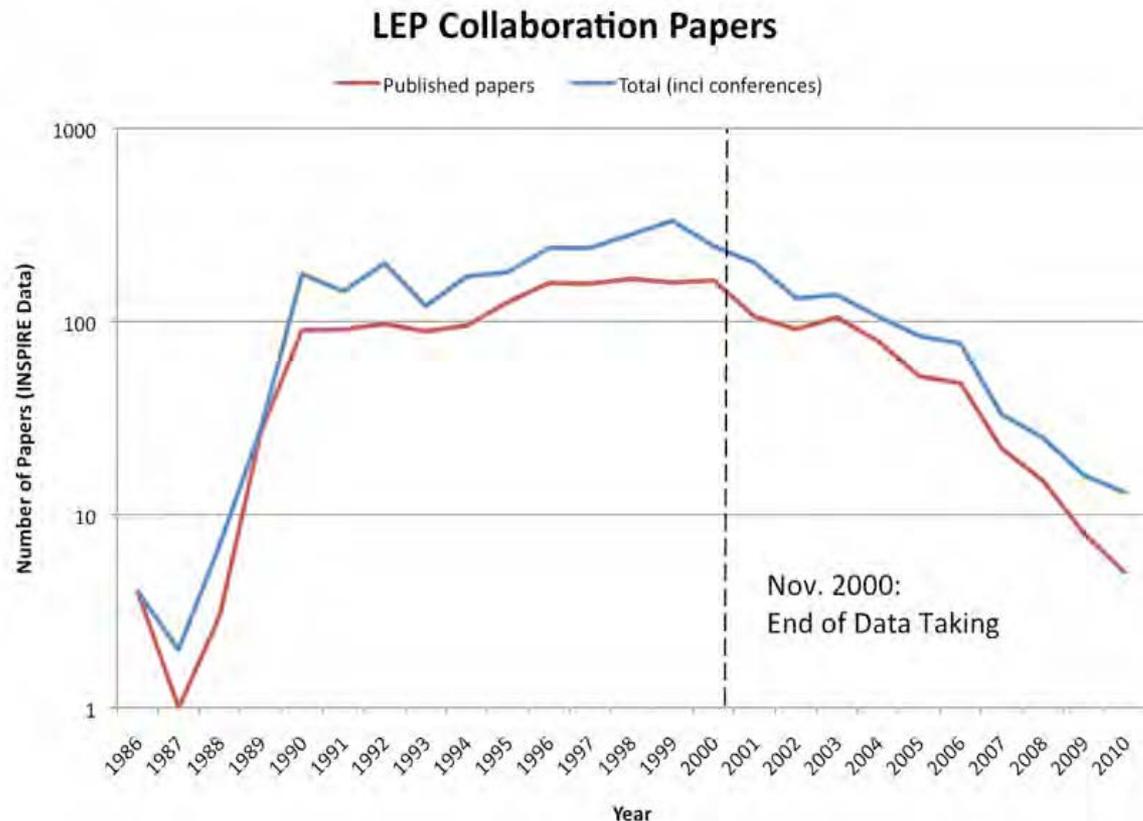
connues : les datations dendrochronologiques précédentes, les analyses à venir sur les restes carbonisés, les données 3D d'un relevé numérique etc. »

### résumé LES DONNÉES

Traitements permettant de rassembler plusieurs banques de données hétérogènes via un modèle commun aux yeux de l'utilisateur.



# Publications à long terme



C. Diaconu, Fredocs, Aussois, 2013

# Echanges : charte

Qui connaît la charte de déontologie des  
métiers de la recherche ?





# Qualité des jeux de données

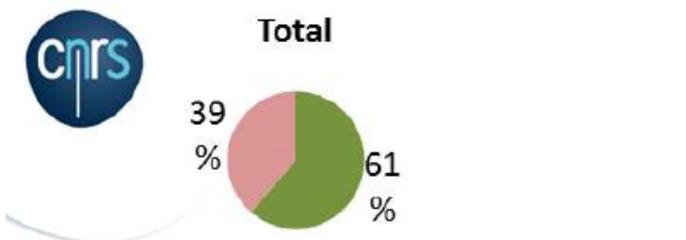
Plusieurs éléments permettent d'évaluer le niveau de qualité d'un jeu de données :

- Des éléments sur les données elles-mêmes et leur structure : format de fichier, structure du fichier, contenu.
- Des éléments attestant du potentiel de réutilisation et de croisement des données : le respect de standards, référentiels et schémas déjà établis ;
- Des éléments qui accompagnent les données : une documentation claire et rigoureuse, gestion des versions et des mises à jour des données...

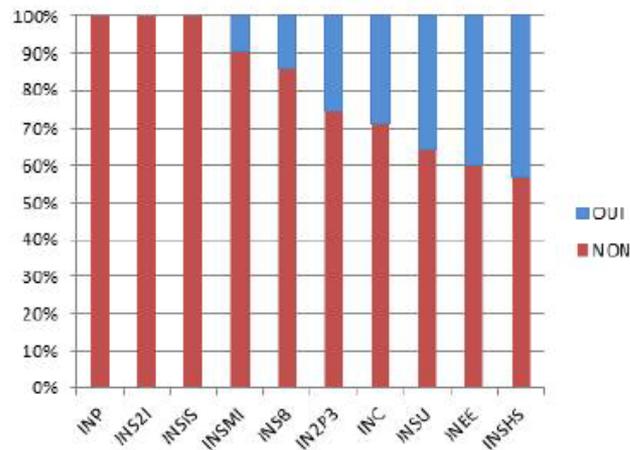
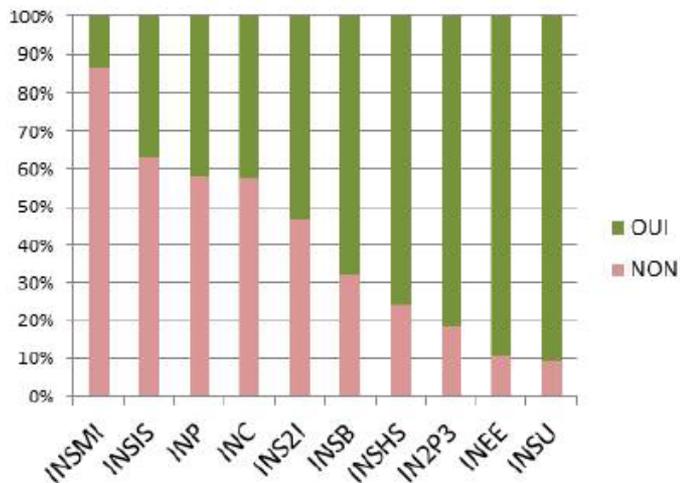


# Enquête DIST (mars 2015)

"PAP 3 47- Les recherches conduites dans votre laboratoire produisent-elles des données de la recherche nécessitant des pratiques de gestion"



Enquête DIST: avez-vous une activité relative aux données de la recherche ?  
OUI:33% NON : 67%





# Atelier données

Une initiative des réseaux...

...accompagnée par la MITI

... complètement en ligne avec le plan données de la  
recherche 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

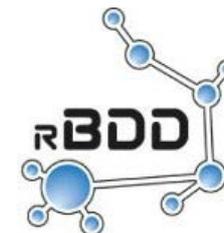
- Apporter une vision transversale de la gestion des données afin d'enrichir la pratique de chaque réseau dans le domaine des données et permettre le développement de la complémentarité entre réseaux.
- Valoriser l'apport des expériences et expertises « métier » dans une vision transversale de gestion de données dans les réseaux technologiques et scientifiques de la MI
- 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.



# Origine

Ce guide est la production du groupe de travail inter-réseaux « Atelier Données » : groupe composé (en 2016) de plusieurs réseaux de la “Mission pour les Initiatives Transverses Interdisciplinaires” (MITI), et de réseaux d’Instituts du CNRS :

- Calcul : réseau pour la communauté du calcul
- Devlog : réseau national des développeurs en logiciel
- Medici : réseau des métiers de l’édition
- QeR : réseau Qualité en Recherche
- rBDD : réseau Bases de données
- Renatis : réseau des professionnels de l’information scientifique
- Resinfo : réseaux des administrateurs systèmes et réseaux
- INIST-CNRS : Institut de l’Information Scientifique et Technique
- SIST : réseau INSU des gestionnaires de données environnementales
- DDOR-CNRS : Direction des données ouvertes de la recherche



# Originalité

L'originalité de ce guide 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.**



# 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



# Le cycle de vie des données

Quelle est votre vision du cycle de vie ?



# Le guide : collecter



Rechercher dans ce livre ...

1. Imaginer et préparer
2. Concevoir et planifier
- 3. Collecter**
4. Traiter
5. Analyser
6. Préserver et archiver
7. Publier et diffuser

Conclusion  
Glossaire  
Infrastructures  
Reproductibilité

## 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 :

[!\[\]\(848edf3a971f9d4a6acd664a9b2a684c\_img.jpg\) 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*

3.1. Utiliser des normes et des standards d'interopérabilité

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

3.2.1. La métrologie des équipements

3.2.2. Les capteurs

3.2.3. Les chaînes de collecte

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

**3.2.5. Les cahiers de laboratoire**

3.2.6. Les tablettes et carnets de terrain

3.2.7. La gestion des collections

3.3. Environnements de stockage - Sauvegarder les données

# Le guide : Préserver et archiver

🔍 Rechercher dans ce livre ...

1. Imaginer et préparer
2. Concevoir et planifier
3. Collecter
4. Traiter
5. Analyser
- 6. Préserver et archiver**
7. Publier et diffuser

Conclusion

Glossaire

Infrastructures

Reproductibilité

Autres guides de bonnes pratiques

Crédits

Document pdf 

[contact](#)

## 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 « Archivaae Numériaue des Données de Recherche ».

LA SCIENCE OUVERTE : vers les connaissances partagées

## 6.1. Comprendre et différencier les différents concepts

### 6.1.1. Définitions générales

- 6.1.2. Préserver la masse de données scientifiques
- 6.1.3. Protéger et sécuriser le patrimoine scientifique

6.2. Préserver les objets numériques

6.3. Archiver les objets numériques

6.4. Sélectionner les données pertinentes

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



# Comment se présente le guide : interactivité et exemples

<https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html>

# 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)*
- *Identification 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

# Stockage / sauvegardes ?



# 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.

# Exemples concrets pour comprendre ces notions

Quels services pour analyser ?



## 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*.

# Pourquoi préserver ? Où et comment ?

Protéger et sécuriser le patrimoine  
scientifique

Sélectionner les données pertinentes



# 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 :

- *le 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 “*Dataper*” pour valoriser et expliciter en détail les données

# Publier les données dans un entrepôt

# Compétences transverses

La gestion des données de la recherche  
nécessite différentes compétences métiers

Comment mettre en place ce type de  
collaboration ?

# Avant de conclure

Que manque t'il à ce guide ?

Comment contribuer ?

# Conclusions

- Ce guide vise à améliorer les pratiques de gestion des données de la science pour :
  - garantir l'intégrité scientifique et la traçabilité de la recherche produite,
  - rendre accessible, partager, permettre la reproductibilité et la réutilisation des données de la recherche : données FAIR
- Les réseaux apportent un fort soutien basé sur une expérience de terrain pour atteindre ces objectifs
- Ce guide n'est pas exhaustif puisqu'il est le reflet des sujets abordés dans le cadre des actions des réseaux impliqués dans la rédaction du guide
- Enrichissez le guide des pratiques métiers d'autres réseaux : nous invitons d'autres réseaux, d'autres entités à nous rejoindre et participer à la prochaine version ... en apportant leurs pratiques métiers dans le cadre de la gestion des données

# Crédits

## Auteurs

Christine Hadrossek : DDOR

Joanna Janik : DDOR

Maurice Libes : réseau SIST

Violaine Louvet : réseau Calcul

Marie-Claude Quidoiz : 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

Site Web : <https://gt-atelier-donnees.miti.cnrs.fr/>

Contact : [donnees-inter-reseaux@services.cnrs.fr](mailto:donnees-inter-reseaux@services.cnrs.fr)