Science Ouverte: Publications, Données & Logiciels

C. Hugon, C. Cavet, B. Khélifi

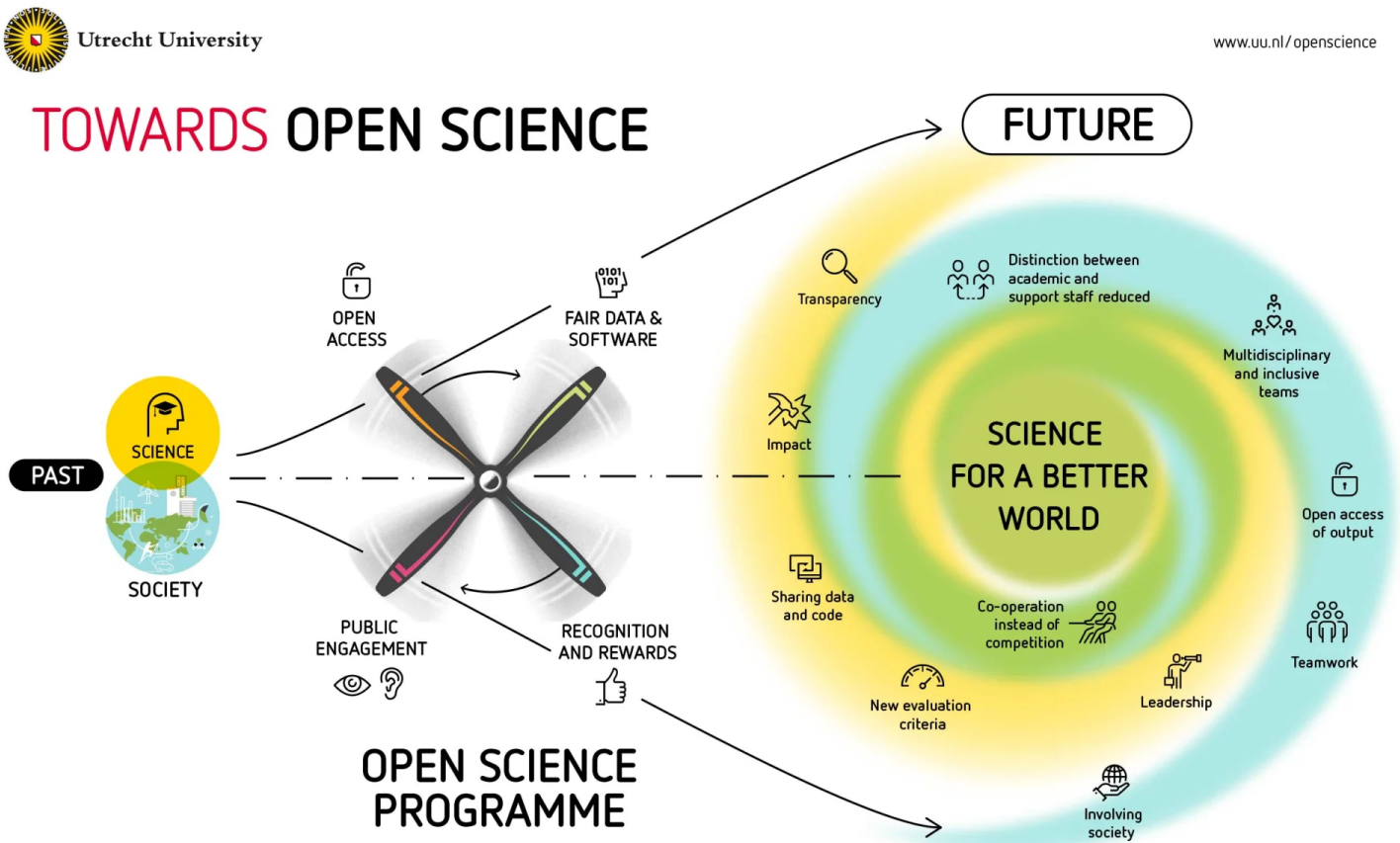
Agenda de l'atelier

- **Présentation des thématiques de la SO**
  - Feuilles de route
  - Les trois piliers
  - Les problématiques transverses
- **Échanges sur les attentes @ APC**

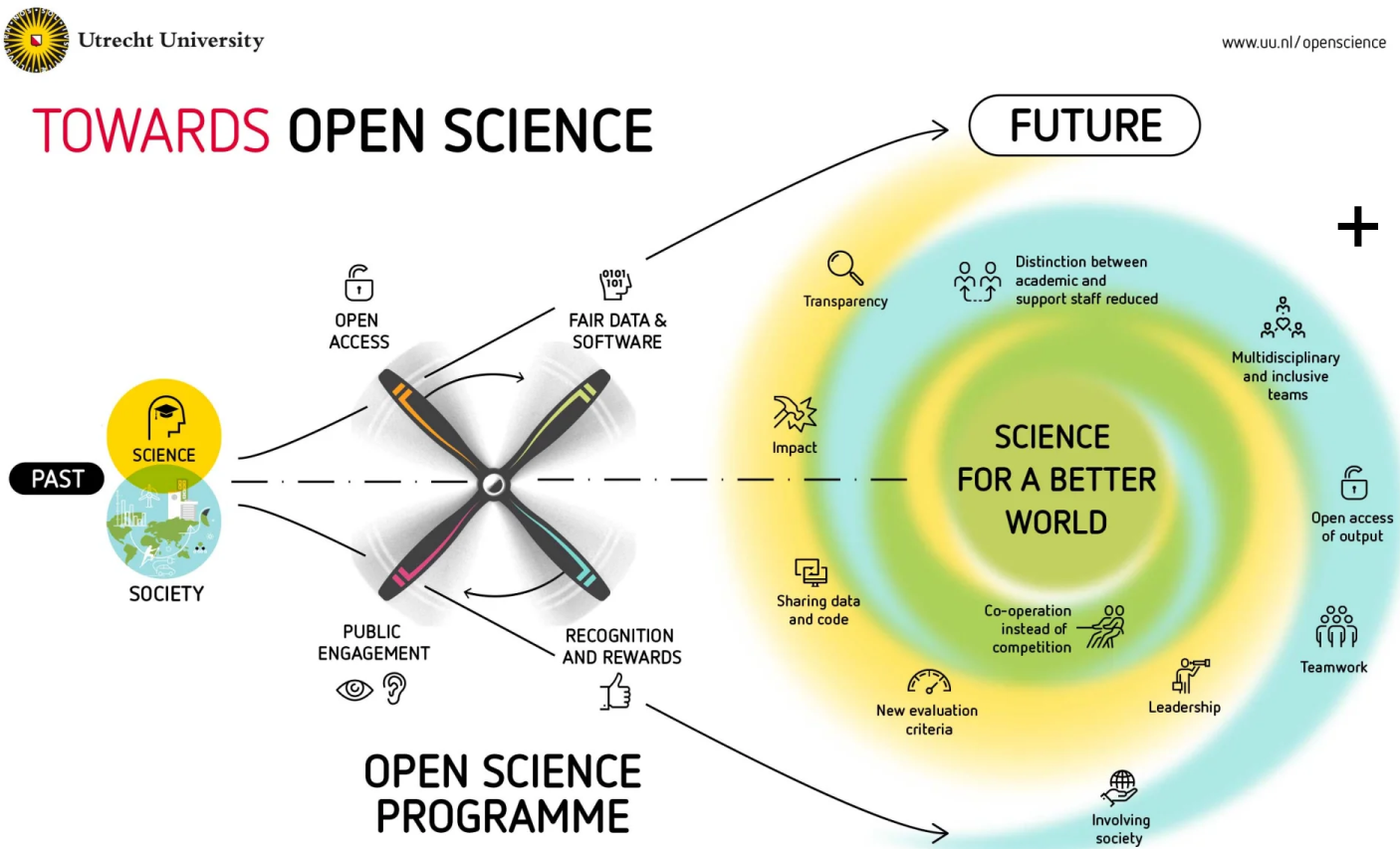
Feuilles de route

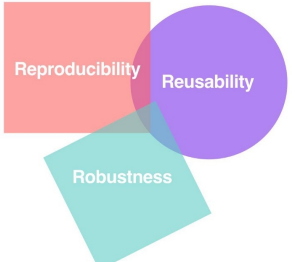
- **Internationales**
  - Texte fondateur 2016 (Amsterdam), UNESCO, Commission Européenne
- **Déclinaisons locales pour la recherche**
  - MESRI : 1er plan en 2018, 2ième en 2021
  - CNRS : 2019
  - Université Paris Cité : 2021
- **Elles sont associées à** :
  - Comité pour la Science Ouverte (CoSO) avec un Fond National pour la Science Ouverte (FNSO)
  - Missions/Groupe de travail au niveau du CNRS (IN2P3 et INSU), de l'UPC, et de l'APC
    - BK : correspondant local SO de l'IN2P3 et l'UPC, Ambassadeur de Software Heritage
    - CH : correspondante locale documentaire de l'IN2P3
    - CC : correspondante locale data de EOSC, sous-comité données UPC

Qu’est-ce Science Ouverte ?

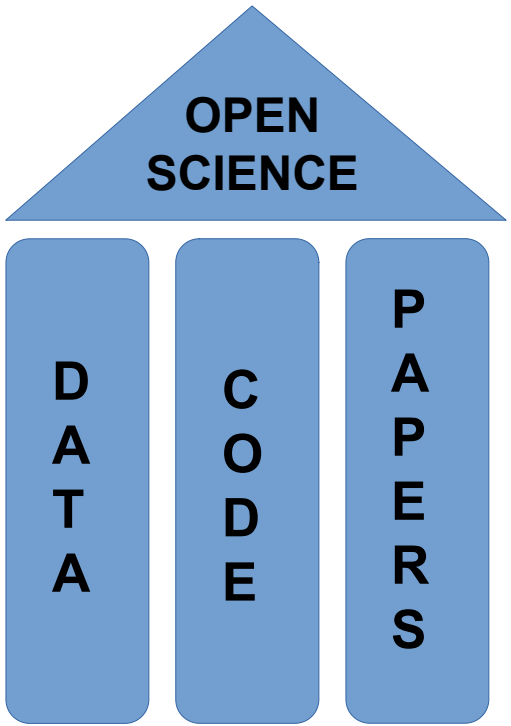


Qu’est-ce Science Ouverte ?

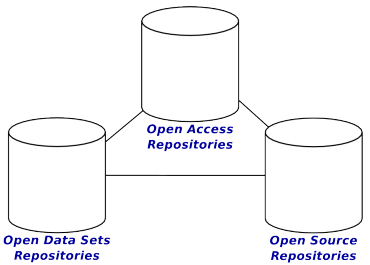




Les trois piliers



- **Infrastructures**



- **Règles**
  - Référencements
  - Principes
  - Etc
- **Méta-données**
  - Contenu, auteurs, etc
  - Provenance
  - Etc

P1 : Publications

Les modèles de publication dans une revue scientifique

- **TRADITIONNEL**
  - Comité éditorial
  - Frais de publication possibles
  - Accès aux publications par abonnements
- **GREEN OPEN ACCESS** (dépôt par l'auteur dans une archive ouverte (HAL/ ArXiv…)
  - Comme preprint
  - Après relecture et publication dépôt du fichier auteur ou éditeur
  - Publication gratuite et consultation accessible à tous

P1 : Publications

Les modèles de publication dans une revue scientifique

- **GOLD OPEN ACCESS**
  - Comité éditorial et relecture par les pairs
  - Publication payées par l'auteur (Article Processing Charges)
  - Lecture sans frais pour tous
- **MODÈLE HYBRIDE**
  - Mélange de modèle traditionnel et de Gold open Access
  - On paye deux fois !

P1 : Publications

Ce que dit la loi

- **100% des textes de publications doivent être déposés dans HAL**

  → Plan National pour la Science Ouverte (PNSO 2018) & Feuille de route CNRS (2019)

- **Une revue ne peut pas refuser à un auteur français le droit de mettre en ligne une version auteur finale acceptée pour publication**

  → Loi Lemaire pour une république numérique (LRN 2016)

- **L'injonction à l'Open Access ne veut pas dire obligation de publier en Gold Open Access**

P1 : Publications

Non au ‘Article Processing Charges’ (APC) !

- Le CNRS encourage ses scientifiques à ne plus payer pour être publiés
- *Même si un contrat de recherche, par exemple issu de l’Agence nationale de la recherche (ANR) ou de l’Europe, permet parfois d’utiliser le financement pour payer des APC, le CNRS demande instamment à ses chercheurs et à ses chercheuses de ne surtout pas payer pour publier un article dans une de ces revues. Ce serait payer deux fois.*
- *Le CNRS demande à celles et ceux qui publient dans une revue sous abonnement de déposer dès sa parution le manuscrit auteur accepté (MAA).*

P1 : Publications

L'OA @ IN2P3 : Processus d'import depuis Inspire vers HAL

- **Inspire moissonne arXiv et les éditeurs**
  - si un arXiv est publié, la notice est mise à jour sans créer de doublon ;
  - si l'arXiv n'est pas publié, il reste dans cet état ;
  - si un article est publié sans arXiv, il est créé à ce moment ;
  - le tout est contrôlé et envoyé à HAL qui fait aussi la mise à jour de statut de notice le cas échéant. HAL fonctionne mal avec les collaborations avec quelques échecs d'import.
- **Pas de différence pour les actes de conférence.**
- **Pas de prise en charge des thèses, livres et chapitres de livres.**

P1 : Publications

UPC : HALathon – CNRS: Sprint CasuHAL

- **Incitation à déposer les textes intégral dans HAL**
- **Bonnes pratiques pour déposer le texte (version auteur ou éditeur)**

ÉVITER LA CRÉATION DE DOUBLONS

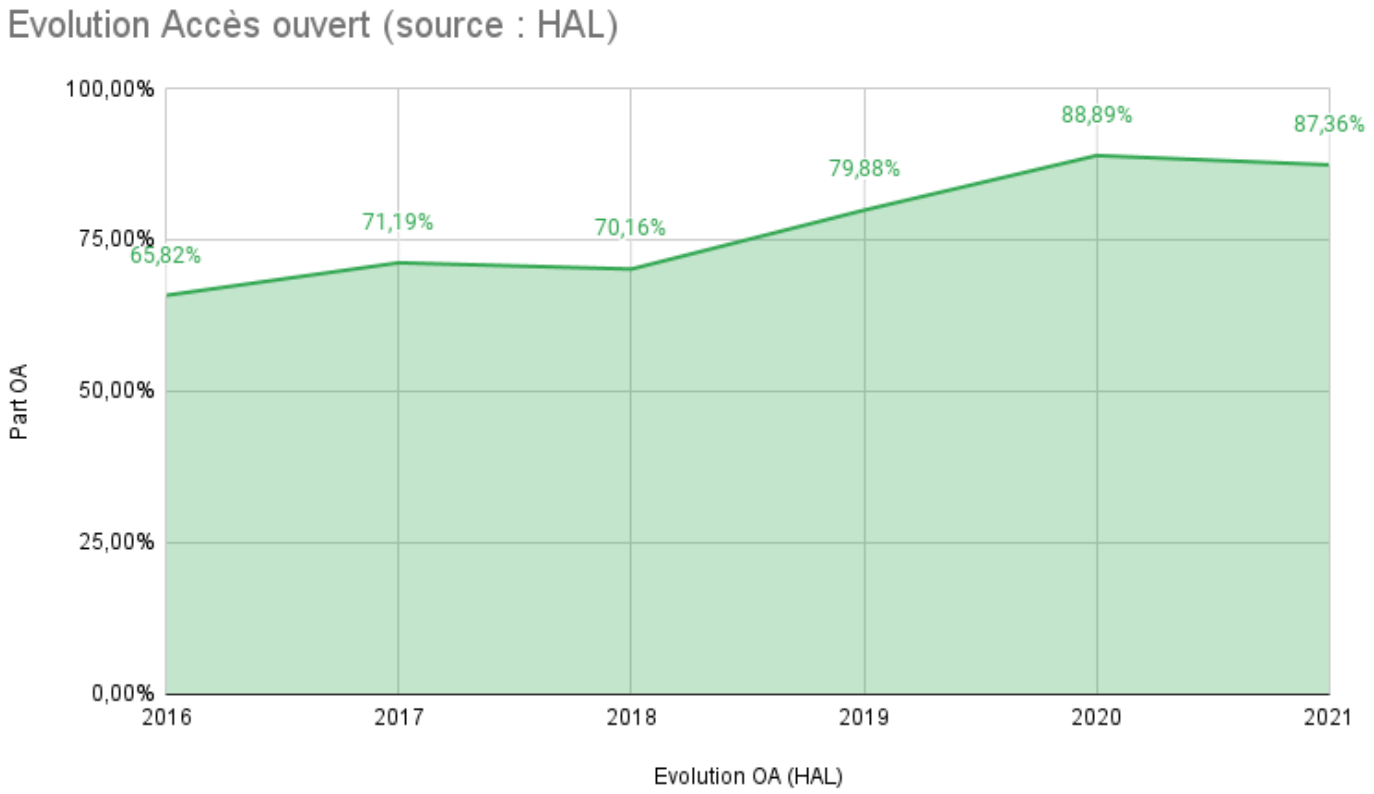
Rechercher la notice dans HAL (recherche avec DOI)

Si elle n'existe pas me le signaler

Si elle existe ajouter le fichier

P1 : Publications

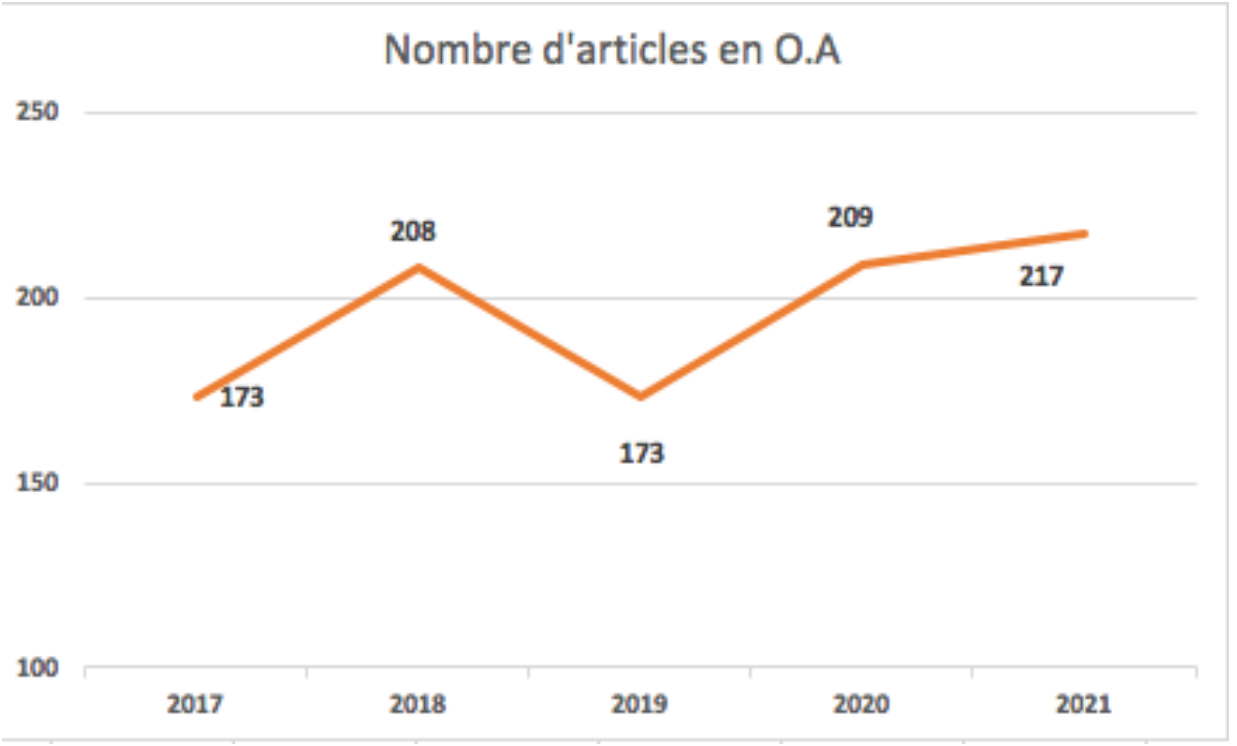
O.A. @IN2P3: En quelques chiffres



P1 : Publications

O.A. @APC: En quelques chiffres

- **Dans INSPIRE hep**
  - Depuis 2017, 93% des articles publiés ont 1 preprint déposés dans ArXiv !
- **Même score dans HAL :**



P1 : Publications

Pour l'HCERES

- Identifier les publications absentes dans INSPIRE
  - Caractériser les thématiques concernées
  - Signalement à IST/IN2P3, pour les intégrer dans le processus d'import

- Campagne d'id ORCID

P1 : Publications

Pour conclure

- **Utiliser le bon format d'affiliation** (voir page Intranet APC/Publication/Affiliation)

Université Paris Cité, CNRS, Astroparticule et Cosmologie, F-75013 Paris, France

- **Déposer les preprints dans ArXviv**
- **Si vous déposez le fichier auteur ou éditeur dans HAL**
  - Rechercher la notice dans HAL à partir du DOI,
    - Ajouter le fichier si la notice existe
    - Contacter Catherine si elle n'existe pas dans HAL
- **Créer votre ID ORCID**, si vous ne l'avez pas !

P2 : Données

Le contexte des feuilles de route

- **Objectifs du CNRS :**

Les données (données brutes, textes et documents, codes sources et logiciels) produites par les chercheurs et les chercheuses CNRS ou avec des moyens mis en œuvre par le CNRS doivent être, dans la mesure du possible, rendues accessibles et ré-utilisables selon les principes FAIR pour une consolidation des connaissances essentielle au développement d'une science plus efficiente. « Les données doivent être aussi ouvertes que possibles, et fermées autant que nécessaire ».

- **Objectifs du UPC** :

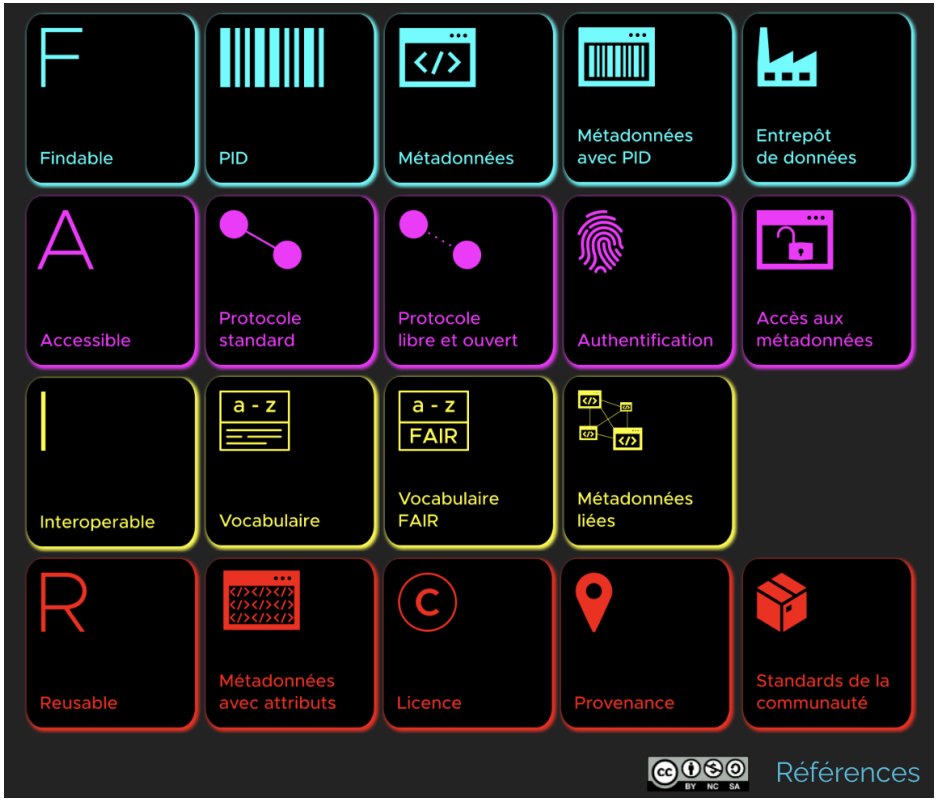
Structurer une politique d'établissement en matière de gestion des données de la recherche en concertation avec ses partenaires nationaux et internationaux, selon le principe de données « aussi ouvertes que possible et aussi fermées que nécessaire » et sur la base de services et outils à valeur ajoutée aussi mutualisés que possible

- **Européen** (lien)
  - Open Data (respectant les principes FAIR)
  - European Open Science Cloud (EOSC)

P2 : Données

Les axes

- **Préparation des données**
  - Curation
  - Data management plan (DMP) / Plan de gestion des données
- **Entrepots de données**
  - Recherche Data Gouv
  - Long term data preservation (LTDP)

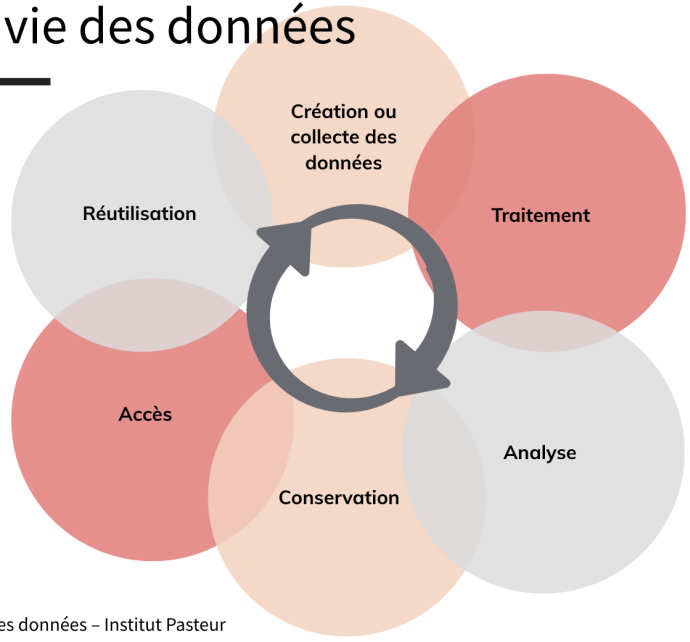


P2 : Données

Préparation des données

- **Curation et documentation des données**
  - Documentation
  - Métadonnées : informations sur la provenance des données, les hypothèses ou contraintes liées à leur production et les protocoles expérimentaux
- **DMP**
  - Gestion des données dans le projet : au cours et à l'issue du projet de recherche
  - Modèles :
    - En fonction des besoins (ANR, Horizon Europe, CCIN2P3, etc.)
    - Outil de rédaction DMP OPIDoR

Le cycle de vie des données



Document adapté de Gérer ses données – Institut Pasteur

Université de Paris

P2 : Données

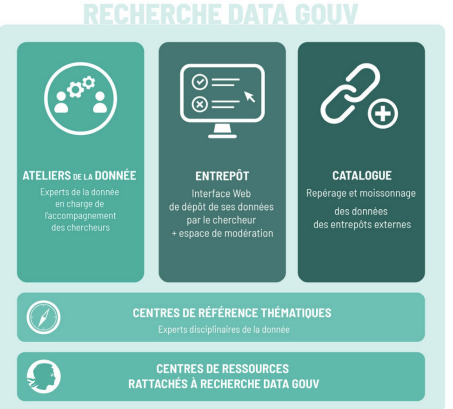
Entrepôts de données

- **But**
  - Identifiant pérenne (Persistent Identifier : PID) de type DOI
  - Description des données : métadonnées descriptives standardisées, vocabulaires disciplinaires contrôlés.
  - Licences et règles d'accès
  - Durée de conservation et archivage pérenne (conservation à moyen et long terme)
- **Entrepôts de données institutionnels, généralistes ou disciplinaires**
  - Pluridisciplinaires et internationaux : Zenodo, Dryad, Figshare
  - Thématiques ou disciplinaires : DataVerse@IPGP
  - Registre d'entrepot de données : **Re3data** (Registry of Research Data Repositories)

P2 : Données

Entrepôts de données nationaux

- L'entrepôt pluridisciplinaire Recherche Data Gouv
  - Opérationnel en 2022
  - Souveraineté française sur les données
  - La pérennité et l'indexation des données stockées, suivant les principes FAIR.

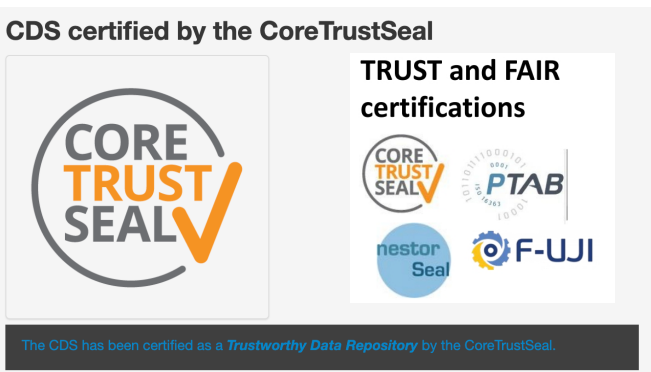


https://projet-recherchedatagv.ouvrirlascience.fr/

P2 : Données

Entrepôts de données européens

- **European Open Science Cloud (EOSC)**
  - partage des données de la recherche et les services associés
- EOSC Association task forces on **LTDP** :
  - **European network of trustworthy digital repositories following FAIR-enabling principles with disciplinary and geographical spread**
  - will provide recommendations for the EOSC board (no implementation)
- **Outils :**
  - Standard
  - Certifications : CoreTrustSeal et FAIR
  - Registre d'entrepot de données : re3data



P3 : Logiciels

- **Objectifs du CNRS :**

Les données (données brutes, textes et documents, codes sources et logiciels) produites par les chercheurs et les chercheuses CNRS ou avec des moyens mis en œuvre par le CNRS doivent être, dans la mesure du possible, rendues accessibles et ré-utilisables selon les principes FAIR pour une consolidation des connaissances essentielle au développement d'une science plus efficiente. « Les données doivent être aussi ouvertes que possibles, et fermées autant que nécessaire ».

- **Objectifs UPC**

Reconnaître les logiciels et les codes, qui sont des sources documentées indispensables à la réutilisation des données de recherche produites, comme des productions scientifiques à part entière et à les intégrer dans sa stratégie de Science ouverte afin qu'ils constituent, avec les publications et les données, le troisième pilier de sa mise en œuvre.

P3 : Logiciels

- Ce pilier est **le moins intégré et unifié**, et pour lequel les réflexions sont les plus intenses
  - Problématiques associées : référencements, inter-référencement, métadonnées, archives, valorisation, évaluation, financement

- Les principes **FAIR4RS** (FAIR for Research Softwares) : Description dans ce lien
  - Comparaison des principes FAIR et FAIR4RS : papier

- **Stockage et archivage en France**
  - HAL utilise officiellement le programme **Software Heritage** pour archiver (sur le long terme) la connaissance associée aux logiciels
  - SW est une archive universelle soutenue par le MESRI (et le CNRS) et le CEA, EOSC, issue du partenariat entre l'INRIA et l'UNESCO en 2017 (lien).
  - Tutoriel de dépôt dans SW/HAL : ici.

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Problématiques transverses

- Évaluation individuelle des chercheurs et enseignants
- Référencement
- Reproductibilité
- Formats de données

Évaluation individuelle des chercheurs et des enseignants

- **Feuille de route européenne**

« The aim is for research to be evaluated based on its intrinsic merits rather than on the number of publications and where these are published. »

- **Objectif CNRS**

Repenser l'évaluation individuelle des chercheurs et des chercheuses avec d'une part l'utilisation d'une évaluation compatible avec les objectifs de la science ouverte et d'autre part la prise en compte de la contribution des chercheurs et des chercheuses à la science ouverte dans l'évaluation

Référencement

- Chaque pilier de la SO devrait avoir son **Identifiant Unique** pour respecter le F des principes FAIR (Findable) ET la notion d'association des Identifiants Uniques entre les 3 produits de la recherche devrait en découler !
  - Données, Logiciel et publications scientifiques sont liés !
  - Les éditeurs, comité d'évaluations, etc ne proposent pas encore les outils adéquats.
  - Les métadonnées associées ne sont pas encore 'universelles' (ex : citation vs authors).
- **Structuration européenne hétéroclite** (Software Heritage, Zenodo, Archives des TGIR et nationales)
  - Multiplication des infrastructures de stockage (Zenodo et le CERN, CDS de Strasbourg , Recherche Data Gouv, etc) qui ne créent pas forcément d'Identifiant Unique: liste non-exhaustive réalisé par EOSC ici.
  - Le financement des infrastructures internationales de la SO est prévu, en France, par le Fond National pour la Science Ouverte (FNSO) : exemple ici

Reproductibilité

- Les dépôts de données de la recherche (brutes, logiciel, texte) n'imposent pas encore tous les éléments pour avoir une réelle reproductibilité sur le long terme (>10 ans)
  - Il manque de nombreuses métadonnées (ex : dépendances, options de compilation).
  - Cela dépend à ce jour de la volonté des auteurs…
- La notion de **Provenance** est intimement couplée à cet objectif
  - Couplage de la provenance du logiciel avec les données pour aboutir à un résultat scientifique.
  - « Provenance as a building block for an Open Science Infrastructure » : lien.
  - Une urgence selon les dernières recommandations de l'UNESCO (lien).
  - Mais comment l'implémenter de manière 'uniforme' ?
    - Groupe de travail dans EOSC (lien), dans l'International Science Council (lien).

Format de données

- Cette problématique est un **point dur technique majeur** pour la Science Ouverte internationale, car le modèle de données n'est pas du tout unifié par thématique.
  - Cela a un impact en profondeur sur les objectifs de chacun des principes FAIR et FAIR2RS :
    - leur référencement (**F**) et leurs métadonnées (**R**),
    - les métadonnées des dépôts de données brutes et de logiciels (**F**),
    - les données elle-même (**I**)
- Ce sujet ne figure pas encore sur les grandes feuilles de route, ne possède pas de référencement comme produit de la recherche et n'est du coup pas/peu valorisé.
  - Mais des organismes et des programmes travaillent dessus (ex : EOSC, ESCAPE, IVOA)

Vos attentes à l'APC

- **Principes de la SO et ses implications 'légales'**
  - Feuilles de route, objectifs à court et long terme, etc
- **La mise en œuvre pratique pour nos recherches**
  - HAL, SW, dépôts de données, etc
- **Attentes localement à l'APC**
  - Documentation, experts/référents, séminaires/hand-on session, etc

Back-up slides….

Les principes FAIR pour les données

→ **Documentation :** lien, pdf

- **F**indable

  The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- **A**ccessible

  Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

- **I**nteroperable

  The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- **R**eusable

  The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Les principes FAIR pour les logiciels

→ **Documentation :** lien, comparaison avec les principes FAIR pour les données

- **F**indable

  The software, and its associated metadata, should be easy to find for both humans and machines.

- **A**ccessible

  The software, and its metadata, must be retrievable via standardized protocols.

- **I**nteroperable

  The software interoperates with other software through exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs).

- **R**eusable

  The software is both usable (it can be executed) and reusable (it can be understood, modified, built upon, or incorporated into other software).