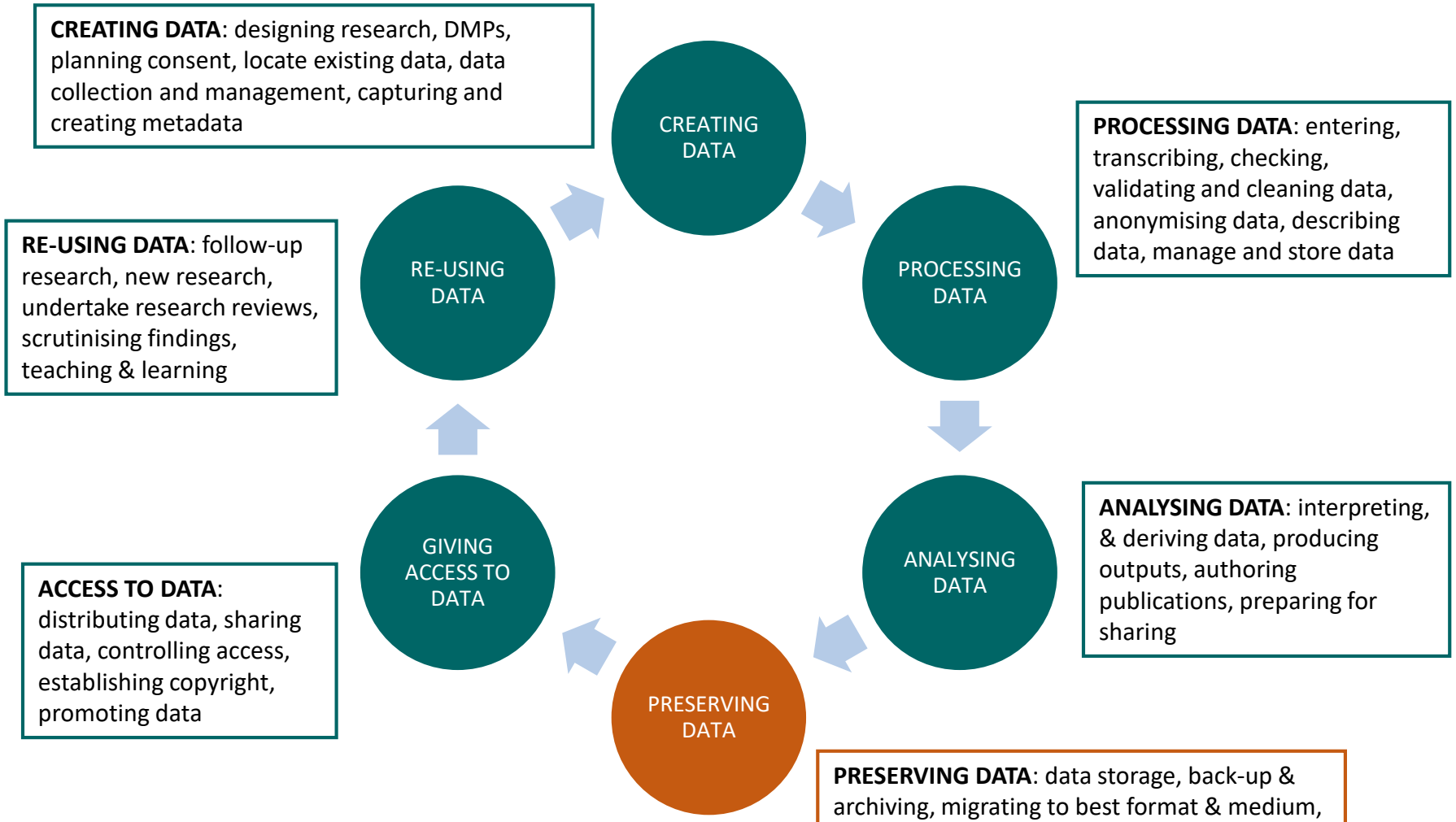
L’archivage pérenne des données scientifiques

Alexia de CASANOVE

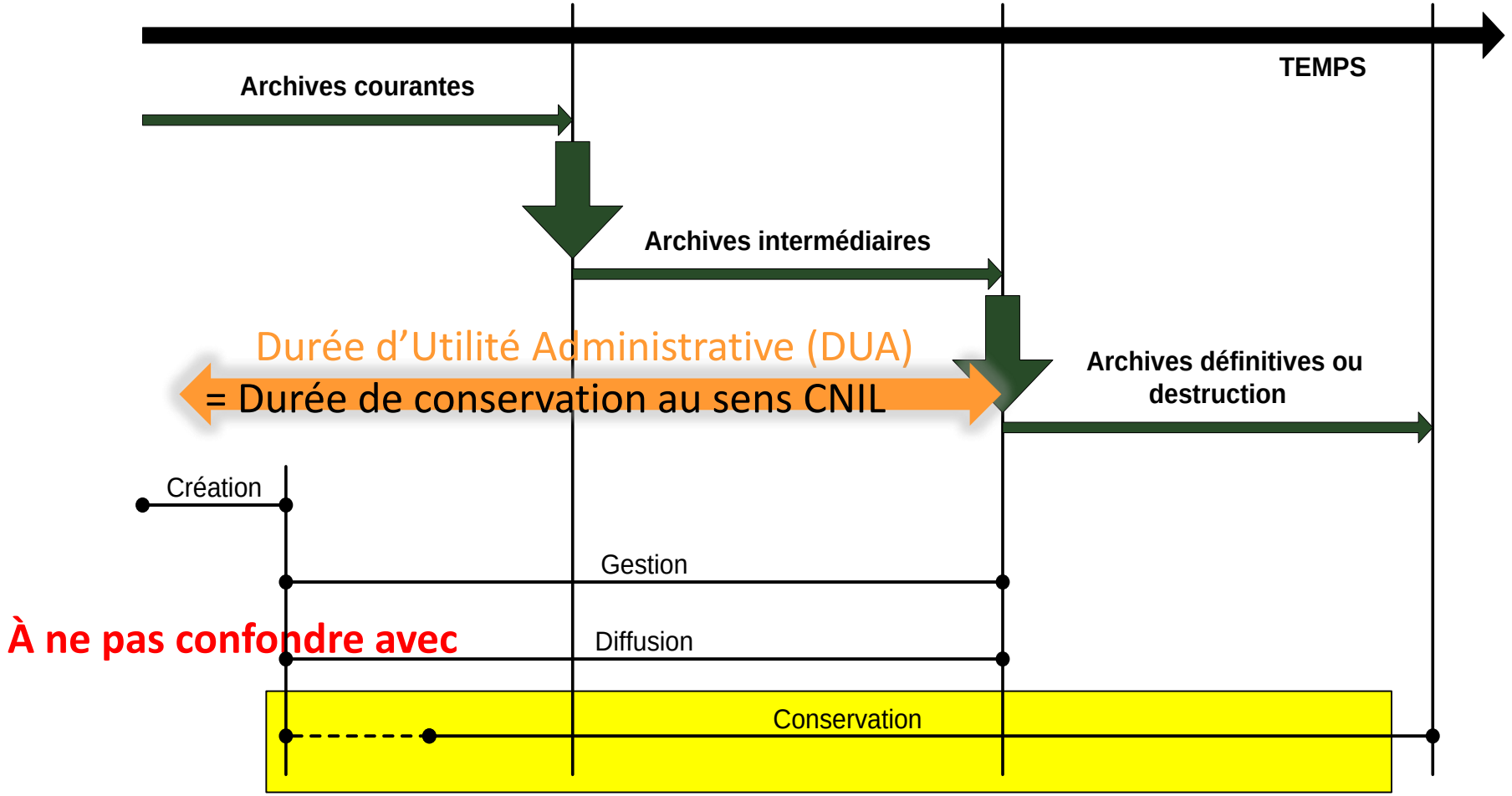
PATC: DMP training – 7 Février 2019

Le cycle de vie des données



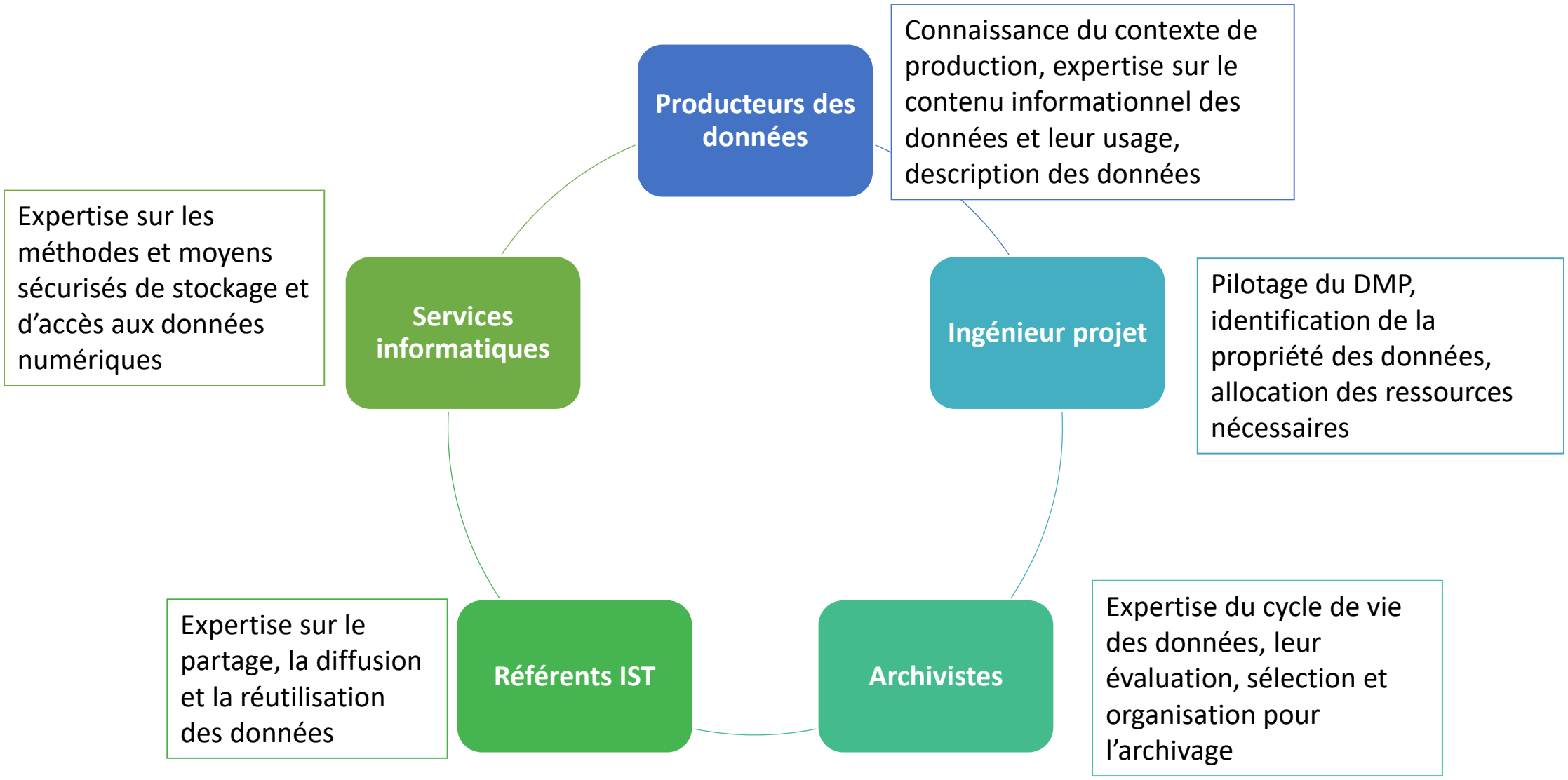
Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

Le cycle de vie des données



À ne pas confondre avec

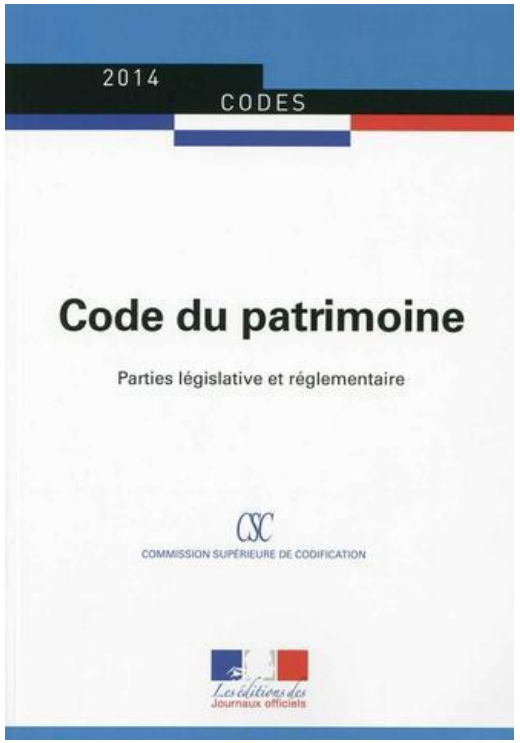
L’archivage dans H2020: les acteurs



L’archivage dans H2020: quelles responsabilités?

- Identifier la propriété des données produites
  - – Déterminée dans l’accord de consortium ?
  - – Porteur de projet français = données relevant de la législation sur les archives publiques ?

- Une archive est publique le plus souvent, car produite :
  - – par des organismes publics,
  - – et/ou dans le cadre d’une mission de service public



Une donnée publique n’est pas forcément publique !

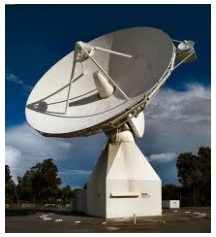
Obligations légales pour les archives intermédiaires

- Gestion à la charge des producteurs / administrations productrices
  - Service dédié pour la gestion des archives
  - Doté de moyens suffisants
  - Externalisation possible mais encadrée
- Quelles responsabilités ?
  - Traiter tous les documents produits, y compris les archives électroniques
  - Suivre les prescriptions en termes de classement et de conservation
  - Respect des règles de communicabilité établies par la loi
- Tri à l'issue de leur durée d'utilité administrative (DUA)
  - Collaboration entre producteurs et archivistes
- Archives à éliminer → visa obligatoire de l'administration des archives
- Archives définitives → conservation dans les services d'archives publics

→ Contrôle scientifique et technique effectué notamment par le Service interministériel des archives de France (SIAF)

Typologies de données

Données d'observation



•Temps réel,

•Uniques, impossibles à reproduire

Relevés météo, images satellite, enquêtes sociales, fouilles archéologiques

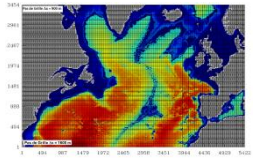
Données expérimentales

•Equipements de laboratoires,

•Reproductibles, parfois coûteuses ou dangereuses

Séquences peptides, poids biomasse, chromatogrammes

Données de simulation numérique



•Modèles informatiques ou statistiques,

•Reproductibles si le modèle est correctement documenté

Modèle climatique, mécanique des fluides.

Données dérivées ou compilées

•Traitement ou combinaison de données « brutes »,

•Reproductibles, mais coûteuses

Base de données compilées, fouille de texte

Données de référence

Séquences de gènes, structures chimiques, etc.

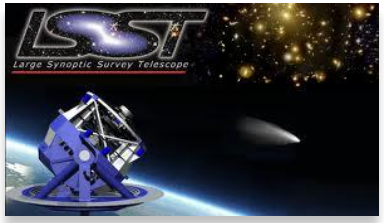
EMBL

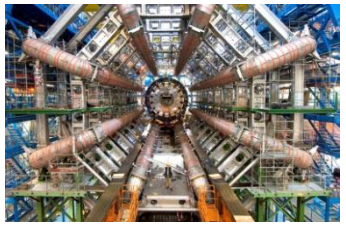
ICGC

(R. Gaillard, 2014; University of Bristol) – Formation URFIST Juin 2015

Panorama de la donnée scientifique numérique

- Démocratisation de la donnée (OpenData)
- Explosion du volume des données:
  - Nouveaux capteurs (plus précis)
    - LSST : Large Synoptic Survey Telescope (15 to 30 Térabytes par nuit).
    - LHC : Large Hadron Collider (Petabytes)
  - Augmentation des capacités de calcul
  - Champs de recherche de plus en plus large





- Exploitation
  - Interdisciplinarité : interdépendance des thématiques scientifiques
  - Data Mining
  - Outils de visualisation et Web 2.0

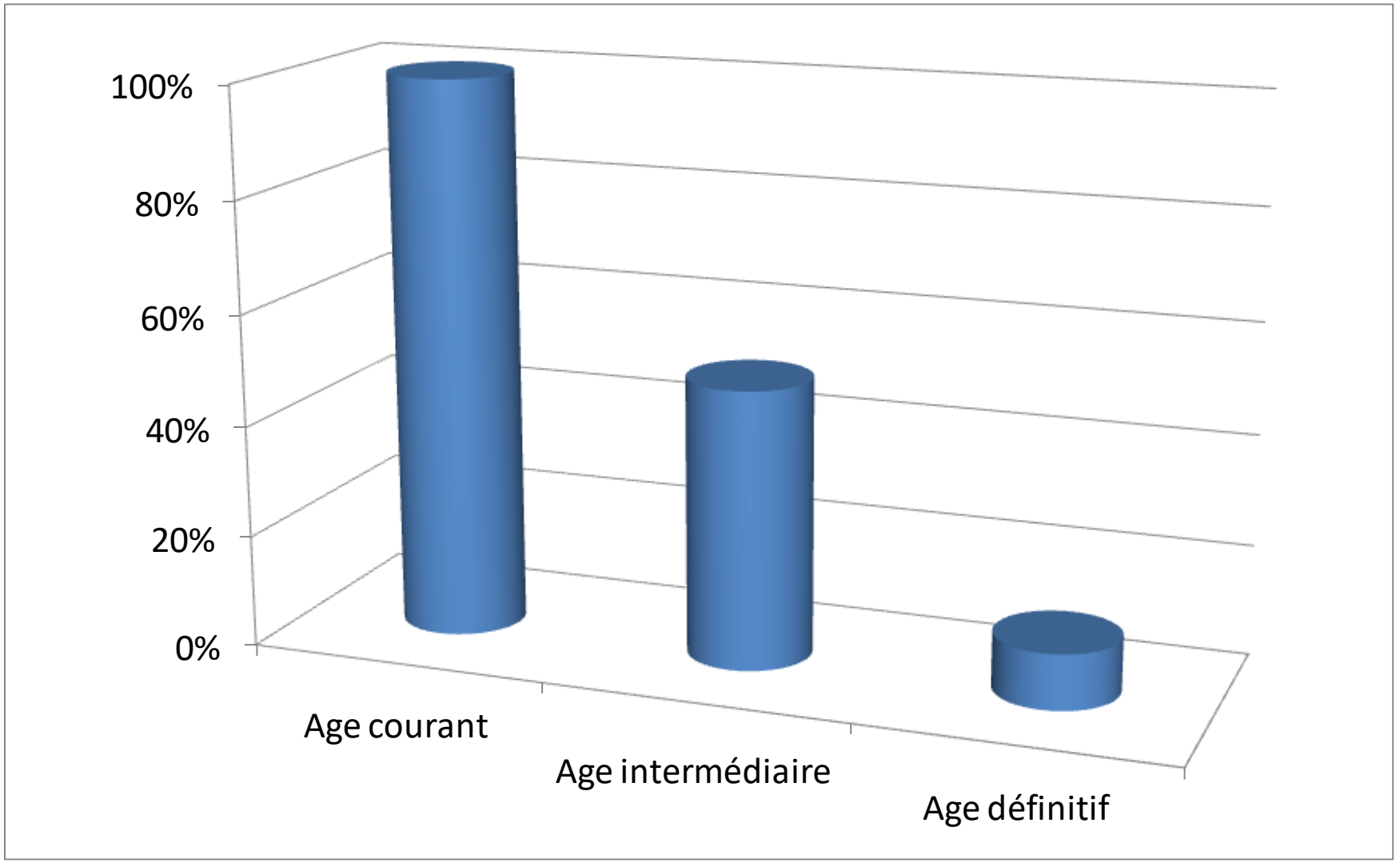
La donnée scientifique est rapidement confrontée aux problématiques du BIG DATA

Panorama de la donnée scientifique numérique

… mais avec des préoccupations supplémentaires dans une perspective d'archivage

- Formats de fichiers complexes et variés
  - Quelques formats « pivot »
    - HDF
    - NetCDF
  - Beaucoup de formats « maison » binaires
- Absence de documentation autour des données
  - Indispensable à la compréhension pour une utilisation future
  - Nécessaire collaboration producteur / archiviste

Quantité de données scientifiques conservées



La problématique de l'archivage numérique

Qu'est-ce que l'archivage électronique ?

L'archivage des documents électroniques consiste à conserver le document et l'information qu'il contient :

- *dans son **aspect physique** comme dans son **aspect intellectuel**,*
- *aussi longtemps que nécessaire (**moyen et long termes**),*
- *de manière à ce qu'il soit en permanence **accessible et compréhensible**.*

Stockage VS Stockage sécurisé VS Archivage

Stockage

Stockage sécurisé

Archivage

Voici un document que j'ai créé en 1998…

De quoi s'agit-il déjà ? Est-ce bien ce qui est indiqué sur la disquette ?

METADONNEES DESCRIPTIVES + IDENTIFICATION
UNIQUE et PERENNE

La disquette est-elle toujours en bon état ?

SUPPORT : VEILLE + MIGRATION PHYSIQUE

Mon portable, acheté en 2017, n'a pas de lecteur de disquette…

ENVIRONNEMENT MATERIEL : VEILLE TECHNO et ANTICIPATION

J'ai créé ce document avec Claris Works. Comment retrouver ce logiciel ? Quel est le format du document ?

ENVIRONNEMENT LOGICIEL : privilégier les FORMATS DURABLES + MIGRATION LOGIQUE

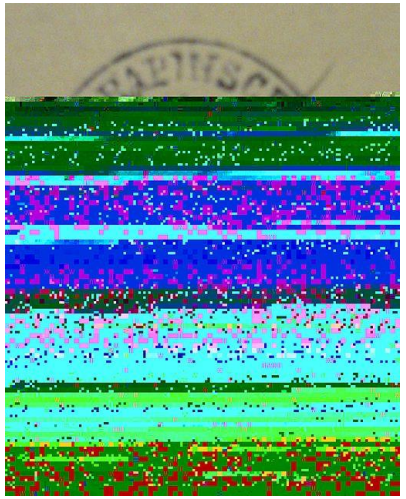
J'ai trouvé le logiciel, mais puis-je l'installer et l'utiliser sous Windows 8 ?

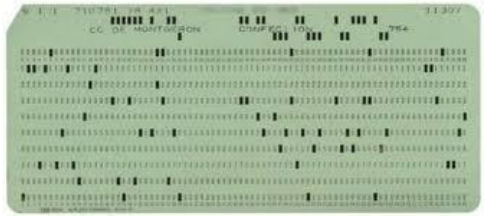
SYSTÈME D'EXPLOITATION

Ça marche ! Mais j'ai perdu toute ma mise en forme…

INTEGRITE AUTHENTICITE

Les enjeux de l’archivage numérique

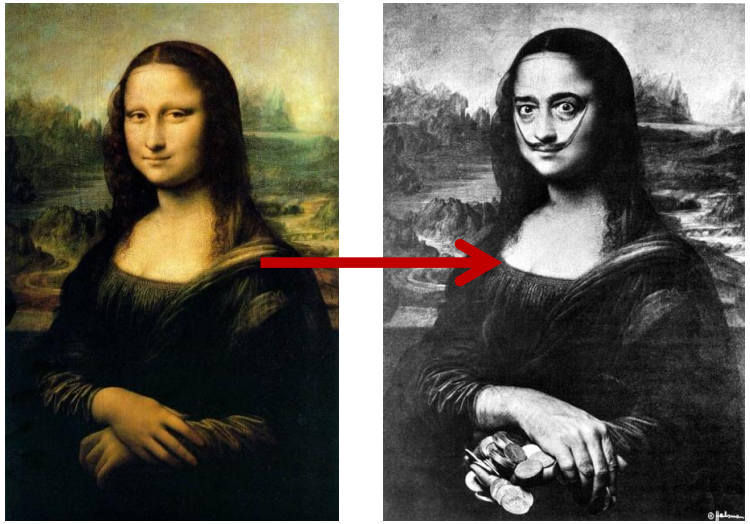




Obsolescence matérielle
Obsolescence logicielle
Obsolescence des formats
Incompréhension « intellectuelle » du contenu
Etc.

Les enjeux de l'archivage numérique

Préservation de l'intégrité



SUPPORT : VEILLE + MIGRATION PHYSIQUE + copie sur SITE DISTANT

Préservation de la lisibilité

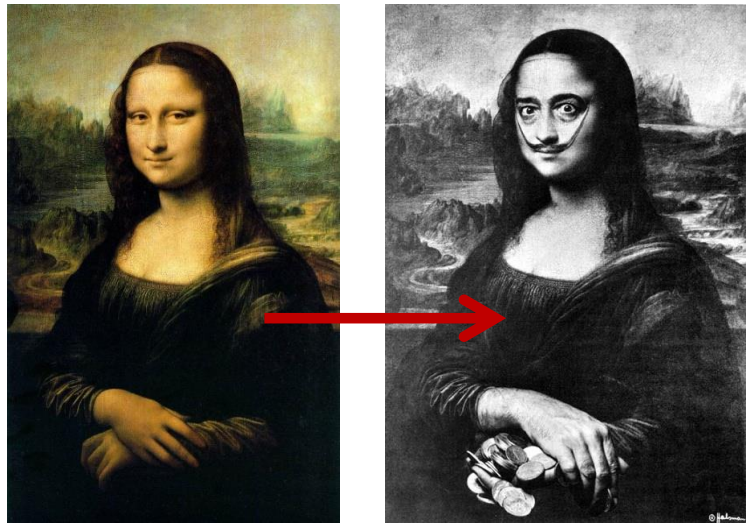
ENVIRONNEMENT MATERIEL : VEILLE TECHNO et ANTICIPATION
ENVIRONNEMENT LOGICIEL : privilégier les FORMATS DURABLES + MIGRATION LOGIQUE (→ validation préalable de format)

Préservation de la capacité à comprendre le contenu du fichier

METADONNEES DESCRIPTIVES + IDENTIFICATION UNIQUE et PERENNE

Intégrité

Comment se protéger de la détérioration des supports?



Comment s’assurer que l’information contenue sur le media n’est pas altérée?

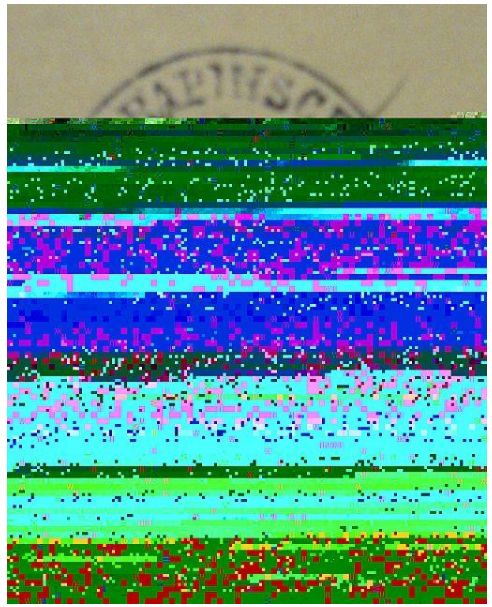
La préservation du train de bits

00100000 00100000 00100000 00110001 00110010 00110000 00100000 00100000
00100000 00100000 00110101 00111000 00100000 00100000 00110101 00000100
00000100 00000110 00100111 00100111 00000110 00100101 10100110 00000111
00100101 10100110 01100110 00001010 10000110 00100110 11101000 11001000
11001000 11101000 11000000 11000100 10111000 11011000 11001000 11100101
01101000 10000000 10000000 10000000 11001100 00011100 10011100 00010111
00011011 10011000 00010000 00010000 00010000 00011001 00011001 10011010
10011001 00010111 00011010 00011010 00000010 00000010 00000010 00000010
00000010 11010011 00010010 11100011 01100011 01100010 00000010 00000010
00000010 11010011 01110010 11011100 11001000 10000000 10000000 10110100
11000100 11100000 10111000 11011000 11000100 10000000 10000000 10110100
11100000 11000000 00110000 00110000 00100000 00100000 00101101 00111000
00110000 00101110 00110000 00110000 00100000 00100000 00100000 00110001
00100000 00000110 10000110 01100100 00000100 00000100 00000111 00100100
00000100 00000100 00000100 00000111 00100111 00000111 00100101 11000110
11001000 10000000 10000000 10000000 10000000 10000000 11000100 11011000
10111000 11000000 11001000 10000000 10000000 10000000 10000000 00011010
10011001 10010111 00011100 00011001 10010000 00000000 10000000 10000000
10000000 11010100 11011000 10111000 11011000 11001000 10000000 10000000
10000000 10000000 10000000 10000000 11010000 10111000 11100000 11000100
10000000 10000000 10000000 10000000 11011100 11100100 11001100 10111000
11010100 11100000 10000000 10000000 10000000 10000000 10000000 10000000
11001100 10111000 11001100 11010000 00101000 10000000 10000000 10000000
11000100 11001000 11000000 10000000 10000000 10000000 10000000 11010100
11100000 10000000 10000000 11010100 10000000 10000000 10000000 11000100
11100100 11100000 11000100 10110100 11000000 11100100 10110100 11001100

MD5 checksum :
a9af0dbc09bab88f9ee2b464142b7b4b

SHA-256 checksum:
33bf8db9db7bd87566584d4095dc1dbb7218e
cd87119e7096d3c8a916f0ddbb9

Algorithmes:
MD5
SHA-256
SHA-512



Lisibilité

Comment accéder à l'information quand le lecteur du média n'existe plus?

Comment ouvrir le fichier quand le logiciel de lecture ne lit plus de fichiers aussi anciens?

Les formats de fichiers

Un format informatique est une convention sur la représentation d'une donnée sur un support numérique. Il peut être :

•**Spécifié** : il existe une description de la convention utilisée pour représenter la donnée et celle-ci est suffisamment décrite pour en développer une implémentation complète.

•**Ouvert** : la convention est publique (sinon le format est dit fermé). Elle est donc sans restriction d'accès ni de mise en œuvre.

•**Normalisé** : la convention est adoptée par des organismes de normalisation (ISO, W3C). Exemple : le PDF/A.

•**Standardisé** : il n'existe pas de norme sur ce format mais son utilisation est tellement répandue qu'il est considéré comme un standard. Exemple : le PDF. ATTENTION : en anglais « standard » signifie « norme ».

•**Propriétaire** : si l'exploitation du format entre dans le cadre du droit privé, il dépend alors de l'existence du propriétaire. Il peut être publié. Exemple : le PDF ou Word.

Ces cinq critères permettent de définir le niveau de pérennité d'un format.

Les formats de fichiers dans le contexte de l'archivage numérique

Une part importante de la problématique de l'archivage pérenne repose sur les formats de fichiers et leur capacité à être interprétés dans un futur lointain.
La condition n° 1 pour qu'un format soit archivable est qu'il doit être exploitable dans son intégralité et sur une durée indéterminée. Pour cela, il doit en exister une spécification accessible qui décrit l'intégralité de ses caractéristiques. Le format et sa spécification doivent être libres de tout droit d'exploitation et sans limite dans le temps.
L'objectif est de trouver un ensemble de critères pour garantir la condition n°1.

Pour être archivé, un format doit donc répondre aux 3 critères suivants :

- Publié
- Largement utilisé (ou promis à l'être)
- **Normalisé (si possible)**

Cette sélection est nécessaire pour :

- le contrôle de la validité d'un format,
- la migration (transformation vers un autre format),
- la lecture et la compréhension du format.

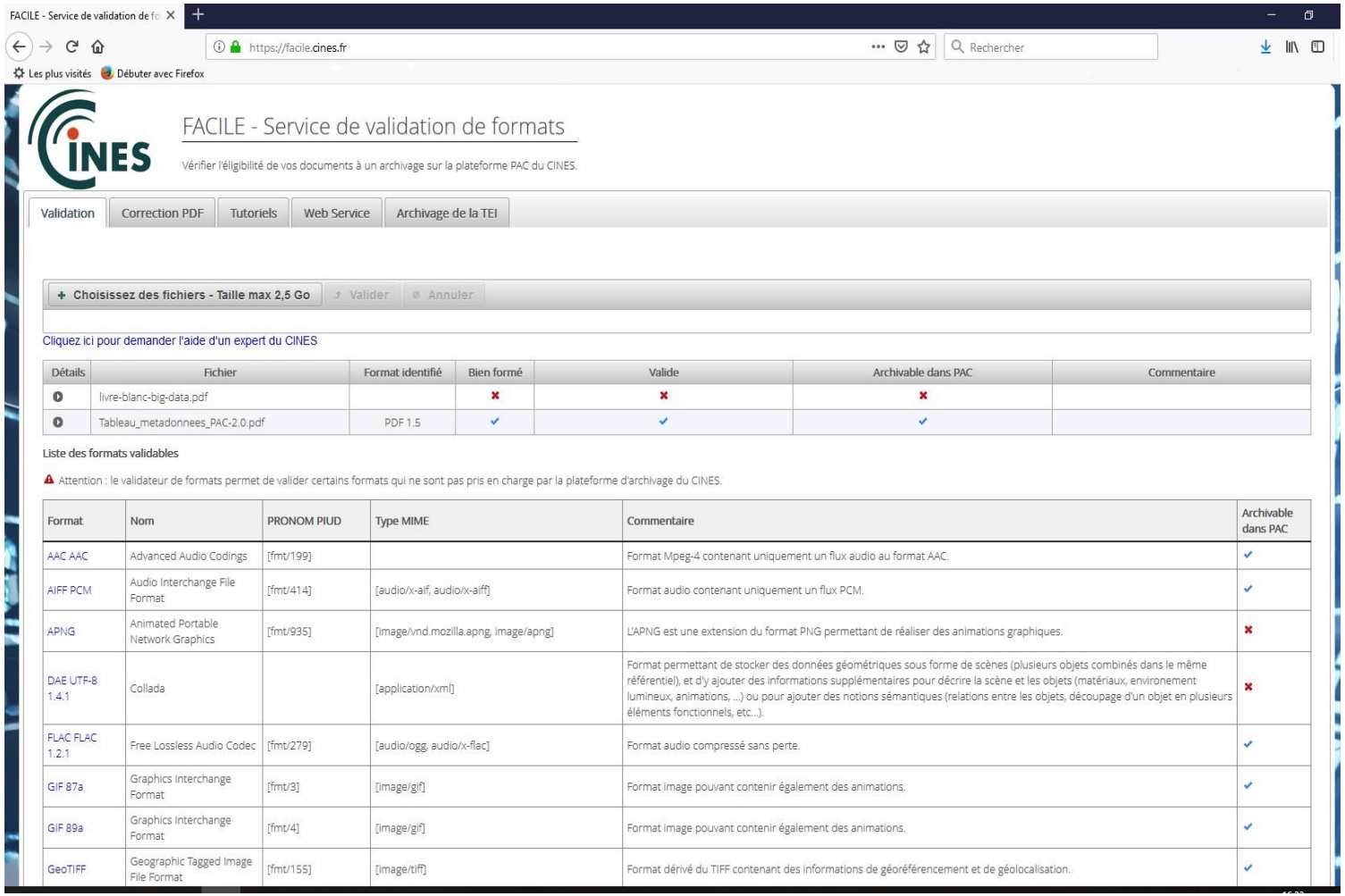
FACILE: Service de validation de formats

Via une interface web, cet outil permet de vérifier si un fichier est valide et bien formé par rapport au format déclaré, et donc de savoir s'il est éligible à l'archivage.

Il suffit pour cela de télécharger le fichier à contrôler. Ce dernier est ensuite analysé par les outils de contrôle qui renvoient automatiquement la réponse.

En cas de fichier mal formé ou non valide, l'utilisateur peut dans un premier temps consulter les tutoriels d'aide à la correction de fichiers disponibles sur l'interface. Si le problème persiste, il peut faire appel à l'expertise du CINES, par mail (analyse de second niveau).

https://facile.cines.fr/

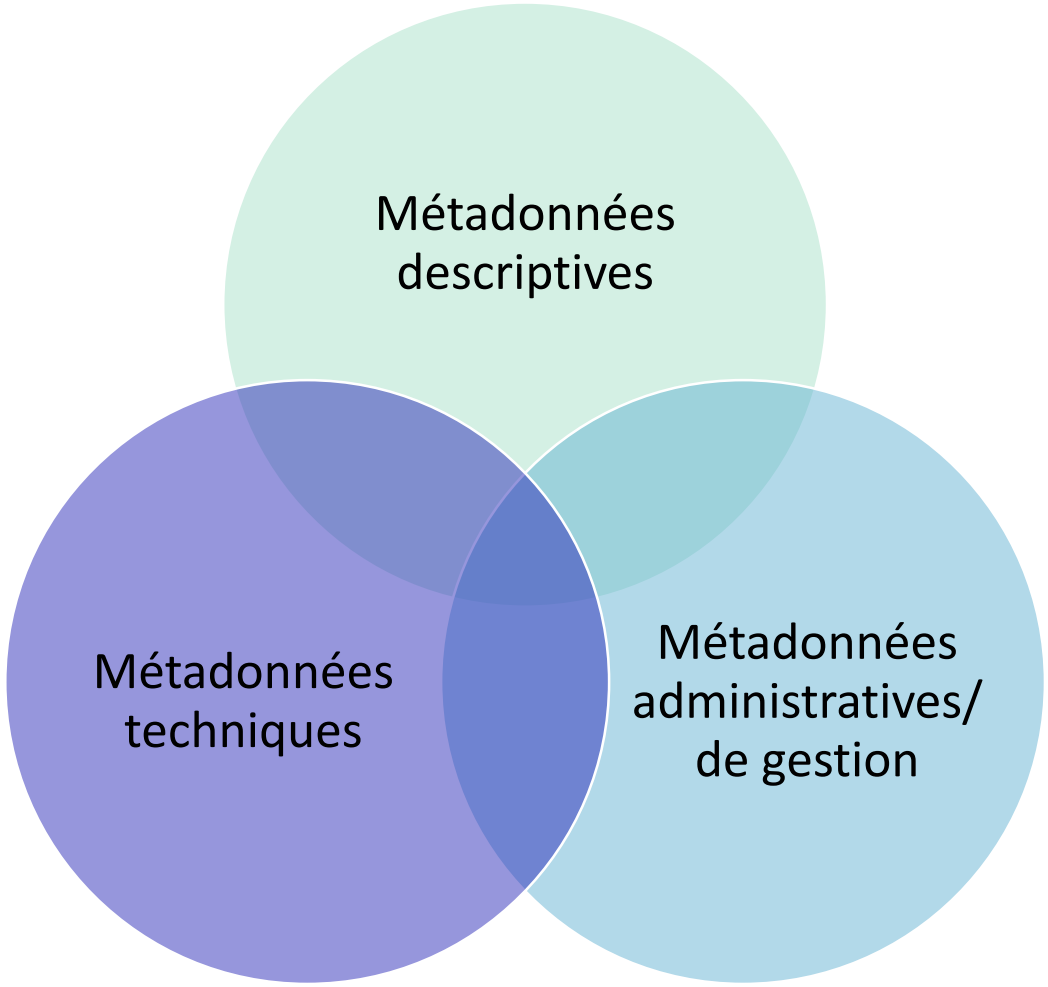


Intelligibilité

Comment contextualiser l’information?

Comment s’affranchir de l’aide du producteur pour comprendre les données?

Les métadonnées pour l'archivage: typologie



Les métadonnées pour l'archivage: typologie

Métadonnées descriptives

• **Identifier un objet ou un groupe d'objets**

- Description bibliographique minimale (ex. : titre, auteur, date, sujet)
- Identifiant unique et pérenne (PID)

• **Appréhender le contenu d'un objet**

- Description bibliographique approfondie et détaillée

• **Identifier les parties qui composent un objet : informations de structure**

- Connaître tous les fichiers qui composent un document
- Connaître la relation physique et logique entre ces fichiers

Métadonnées de gestion / administratives

• **Identifier / authentifier l'auteur** (signature électronique)

• **Gérer le cycle de vie des objets**

- Gestion de la DUA et du sort final
- Migrations et toutes modifications affectant le train de bits de l'archive
- Modifications des métadonnées
- Construire et maintenir des liens entre les versions

→ **Pour conserver l'historique de la création et des modifications subies par l'objet numérique**

• **Garantir l'intégrité d'un fichier**

- Vérification du train de bits par empreinte / checksum

• **Gérer les droits d'accès**

Métadonnées techniques

• **Informations sur les plateformes** (environnement matériel et logiciel) **pour l'émulation**

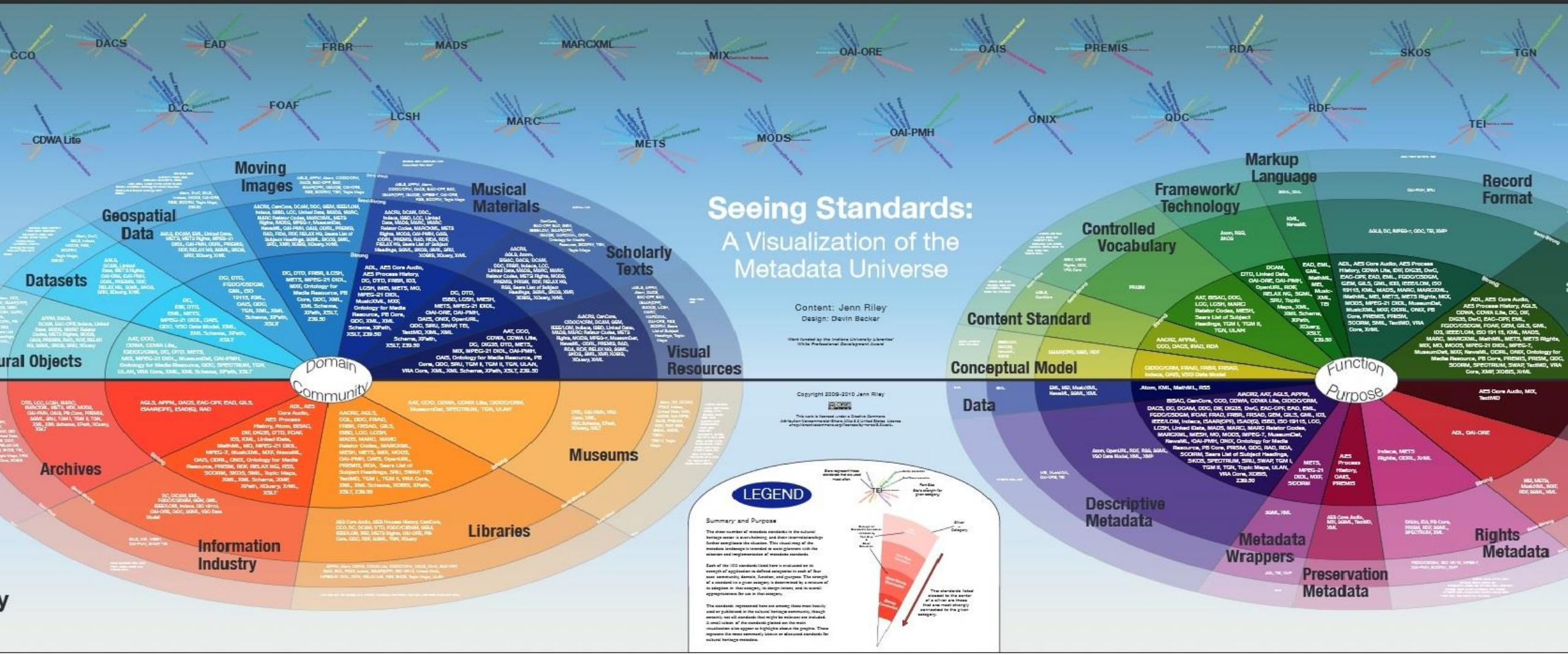
• **Informations sur les fichiers** (taille, compression,…)

• **Informations sur les formats eux-mêmes**

- Répertoires de formats
- Évaluation des formats

• **pour la conservation (migration, émulation)**

• **pour la restitution (pour savoir comment visualiser ce qu'on a conservé)**

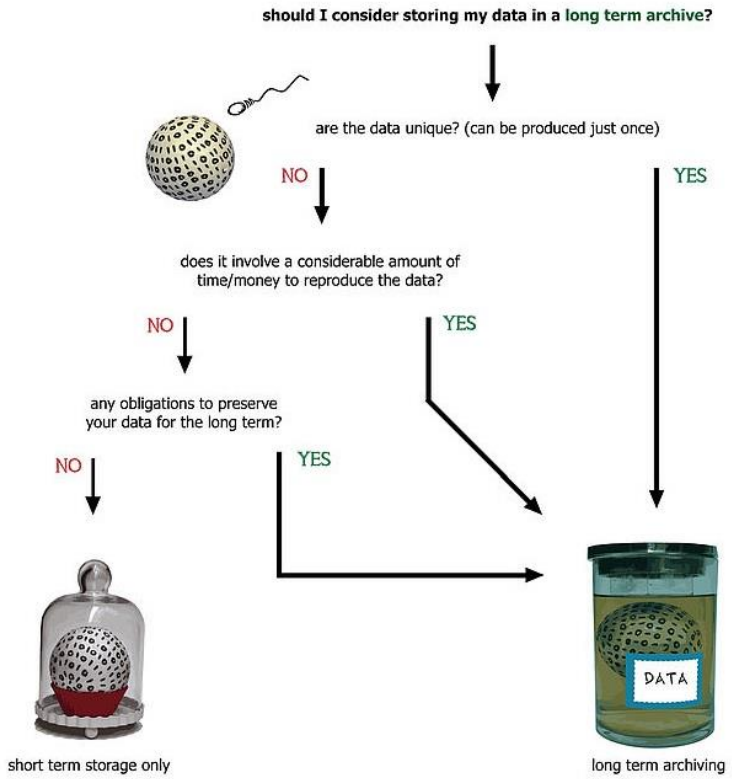


http://jennriley.com/metadatamap/seeingstandards.pdf

Que faut-il archiver?

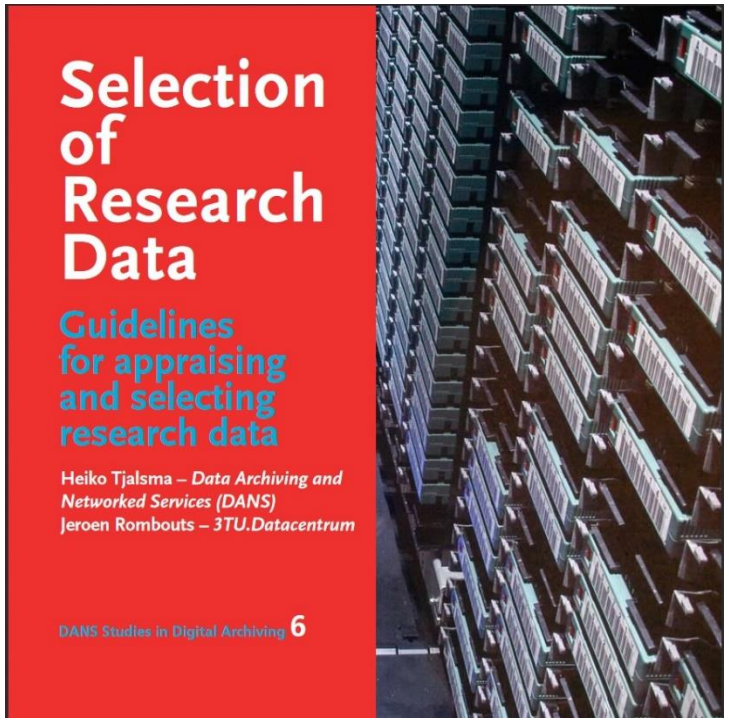
Decision diagram for data selection (Research Data Netherlands):

https://datasupport.researchdata.nl/start-de-cursus/iv-gebruiksfase/data-archiveren/selectie-van-data/



Selection of Research Data, DANS:

https://dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANSselectionofresearchdata.pdf



Que faut-il archiver?

Référentiel de gestion des archives de la recherche:

https://www.archivistes.org/IMG/pdf/referentiel_recherche_intro_septembre2012_corrige_.pdf

Section Aurore

8 rue Jean-Marie Jégo
75013 Paris
Tel : 01.46.06.39.44
Fax : 01.46.06.39.52
secretariat@archivistes.org
www.archivistes.org

Référentiel de gestion des archives de la recherche

Objet

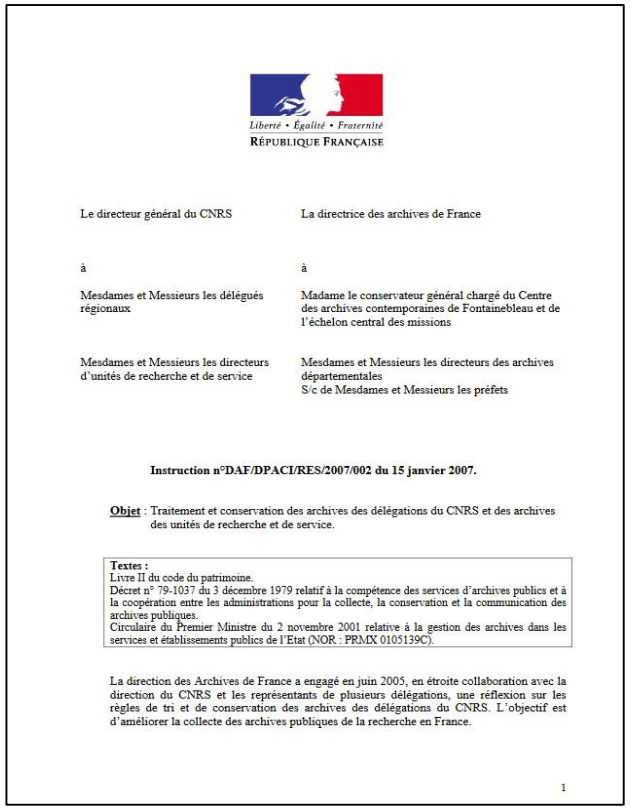
Traitement et conservation des archives des laboratoires de recherche, des chercheurs et des enseignants chercheurs dans les universités, les organismes de recherche et les agences sanitaires.

Textes de référence

- Livre II du Code du patrimoine.
- Décret N° 79-1037 du 3 décembre 1979 modifié relatif à la compétence des services d'archives publics et à la coopération entre les administrations pour la collecte, la conservation et la communication des archives publiques.
- Instruction DPACI/RES/2005/003 du 22 Février 2005 relative au tri et à la conservation des archives reçues et produites par les services et établissements concourant à l'éducation nationale.
- Instruction DAF/DPACI/RES/2007/002 du 15 janvier 2007 relative au traitement et à la conservation des archives des délégations du CNRS et des archives des unités de recherche et de service.
- Circulaire du Premier ministre du 2 novembre 2001 relative à la gestion des archives dans les services et établissements publics de l'État (NOR : PRMX 0105139C).

Instruction sur les archives du CNRS:

https://francearchives.fr/file/f4b709862fb6a7048925ed9ba24cb3181bab6c6a/static_871.pdf



Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Le directeur général du CNRS

à

Mesdames et Messieurs les délégués régionaux

Mesdames et Messieurs les directeurs d'unités de recherche et de service

La directrice des archives de France

à

Madame le conservateur général chargé du Centre des archives contemporaines de Fontainebleau et de l'échelon central des missions

Mesdames et Messieurs les directeurs des archives départementales
S/c de Mesdames et Messieurs les préfets

Instruction n°DAF/DPACI/RES/2007/002 du 15 janvier 2007.

Objet : Traitement et conservation des archives des délégations du CNRS et des archives des unités de recherche et de service.

Textes :
Livre II du code du patrimoine.
Décret n° 79-1037 du 3 décembre 1979 relatif à la compétence des services d'archives publics et à la coopération entre les administrations pour la collecte, la conservation et la communication des archives publiques.
Circulaire du Premier Ministre du 2 novembre 2001 relative à la gestion des archives dans les services et établissements publics de l'Etat (NOR : PRMX 0105139C).

La direction des Archives de France a engagé en juin 2005, en étroite collaboration avec la direction du CNRS et les représentants de plusieurs délégations, une réflexion sur les règles de tri et de conservation des archives des délégations du CNRS. L'objectif est d'améliorer la collecte des archives publiques de la recherche en France.

1

Que faut-il archiver?

Sensibilisation à la sécurisation et à la pérennisation des données, 6 novembre 2014

https://webcast.in2p3.fr/container/rbdd2014

- Retour d'expérience : les données de l'Ecothèque méditerranéenne (CEFE)
- Quels sont les critères à prendre en compte pour la conservation des données (CINES)
- Une préoccupation partagée : plan de gestion des données et projets Horizon 2020 (France Grilles)
- Méthodes développées par le centre de données astronomiques de Strasbourg
- La pérennisation des données chiffrées ? Quel est l'impact du chiffrement sur le long terme ?
- Retour d'expérience sur l'utilisation du format SIARD pour l'archivage des bases de données relationnelles
- Organisation du Centre de ressources numériques Cocoon
- Table ronde et synthèse et perspectives au sein du réseau Base de données

Focus: les “Trustworthy Digital Repositories”

Trustworthy Digital Repository = plateforme d’archivage certifiées

- **Core Trust Seal** (https://www.coretrustseal.org/

ex- DSA, https://www.datasealofapproval.org/en/)

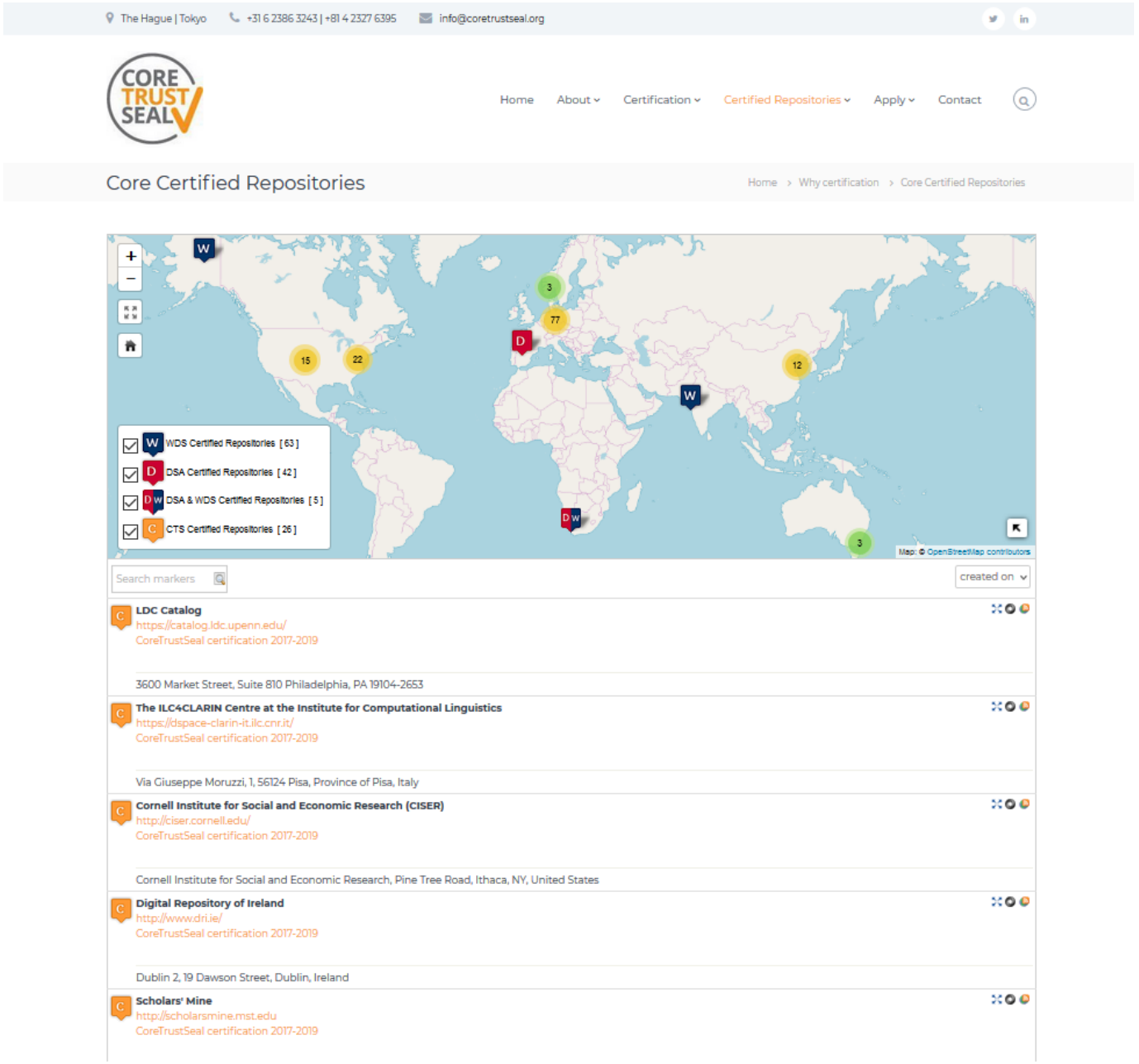
- **Nestor** (http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html)
- **ISO 16363** (https://www.iso.org/standard/56510.html)

Ces certifications font partie de l’« European Framework for Audit and Certification of Digital Repositories »

Elles se concentrent sur l’organisation structurelle, la gestion des objets numériques et les technologies employées.

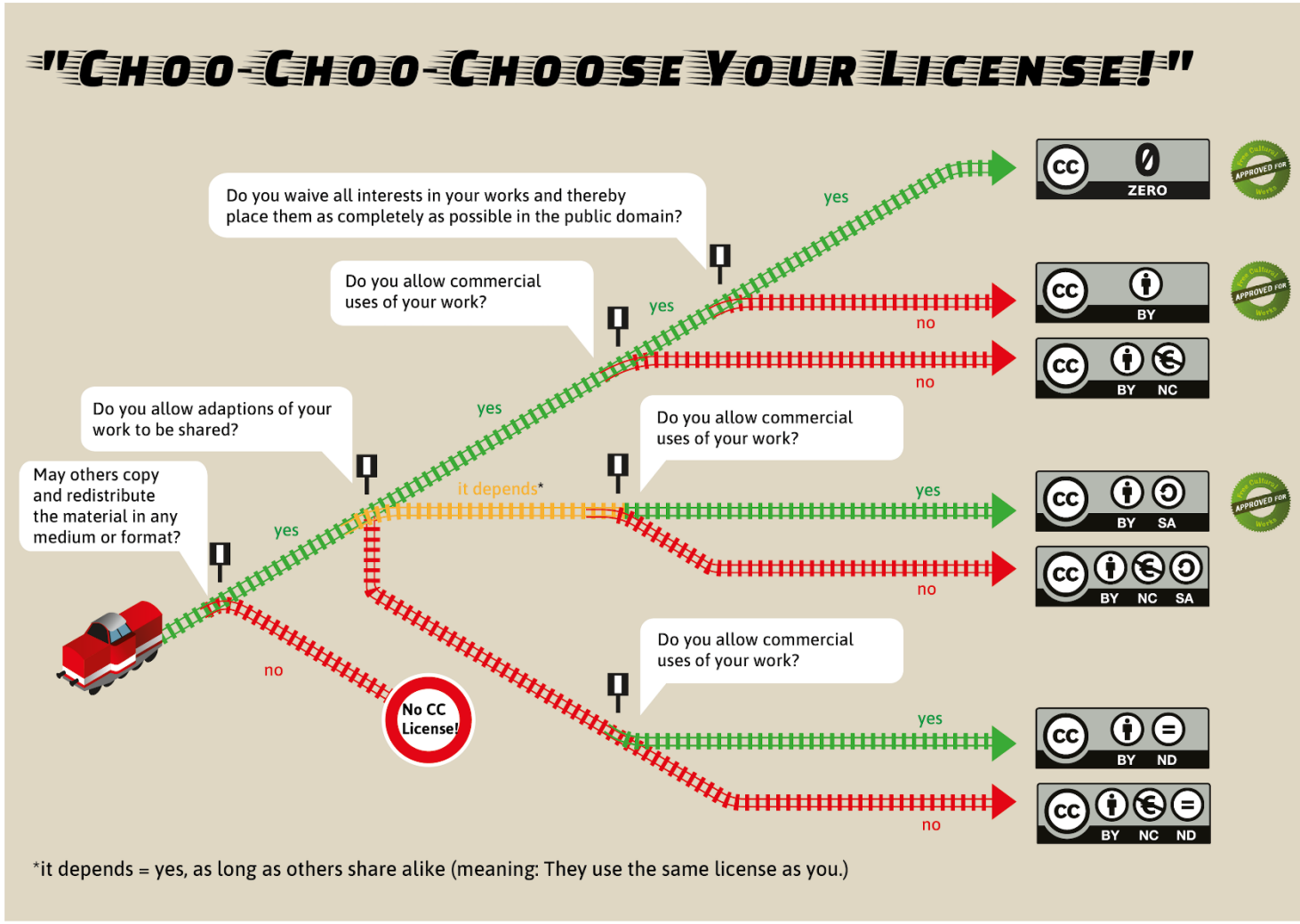
Modèle OAIS:

Les 6 fonctions du modèle OAIS

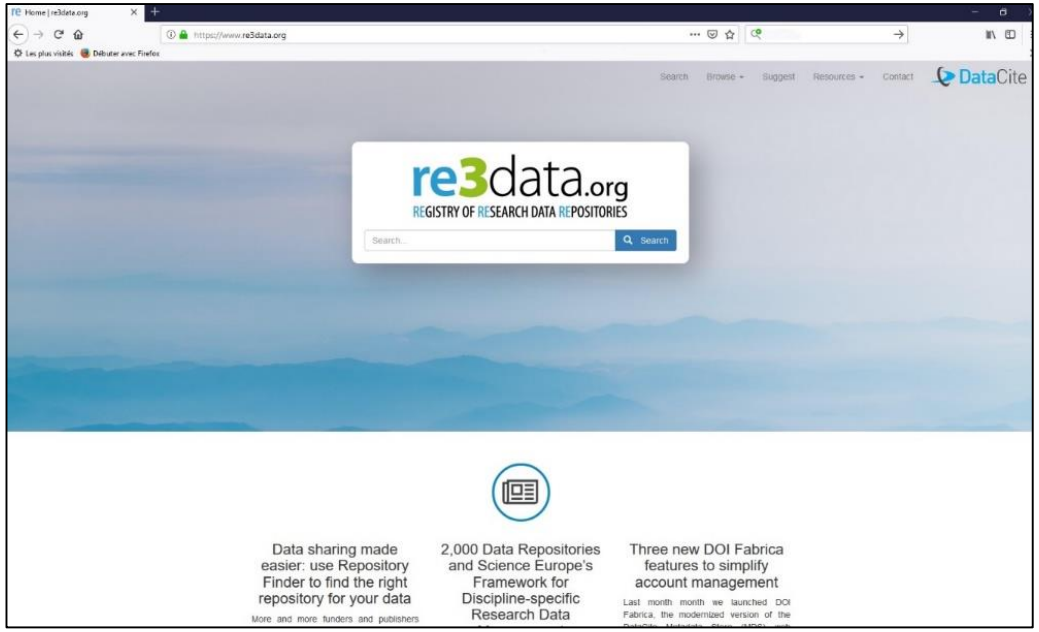

The Hague | Tokyo
+31 6 2386 3243 | +81 4 2327 6395
info@coretrustseal.org
CORE TRUST SEAL
Home
About
Certification
Certified Repositories
Apply
Contact
Core Certified Repositories
Home › Why certification › Core Certified Repositories
WDS Certified Repositories [63]
DSA Certified Repositories [42]
DSA & WDS Certified Repositories [5]
CTS Certified Repositories [26]
Search markers
created on
LDC Catalog
https://catalog.ldc.upenn.edu/
CoreTrustSeal certification 2017-2019
3600 Market Street, Suite 810 Philadelphia, PA 19104-2653
The ILC4CLARIN Centre at the Institute for Computational Linguistics
https://dspace-clarin-it.ilc.cnr.it/
CoreTrustSeal certification 2017-2019
Via Giuseppe Moruzzi, 1, 56124 Pisa, Province of Pisa, Italy
Cornell Institute for Social and Economic Research (CISER)
http://ciser.cornell.edu/
CoreTrustSeal certification 2017-2019
Cornell Institute for Social and Economic Research, Pine Tree Road, Ithaca, NY, United States
Digital Repository of Ireland
http://www.dri.ie/
CoreTrustSeal certification 2017-2019
Dublin 2, 19 Dawson Street, Dublin, Ireland
Scholars' Mine
http://scholarsmine.mst.edu
CoreTrustSeal certification 2017-2019

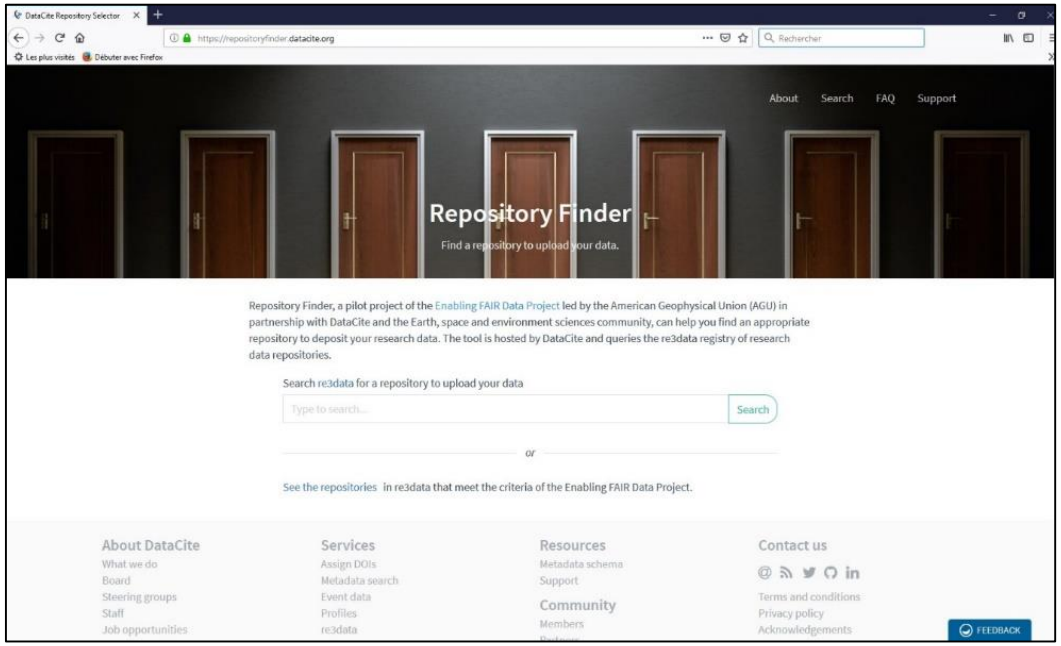
ANNEXES

Licences



Entrepôts de données





https://repositoryfinder.datacite.org/

zenodo

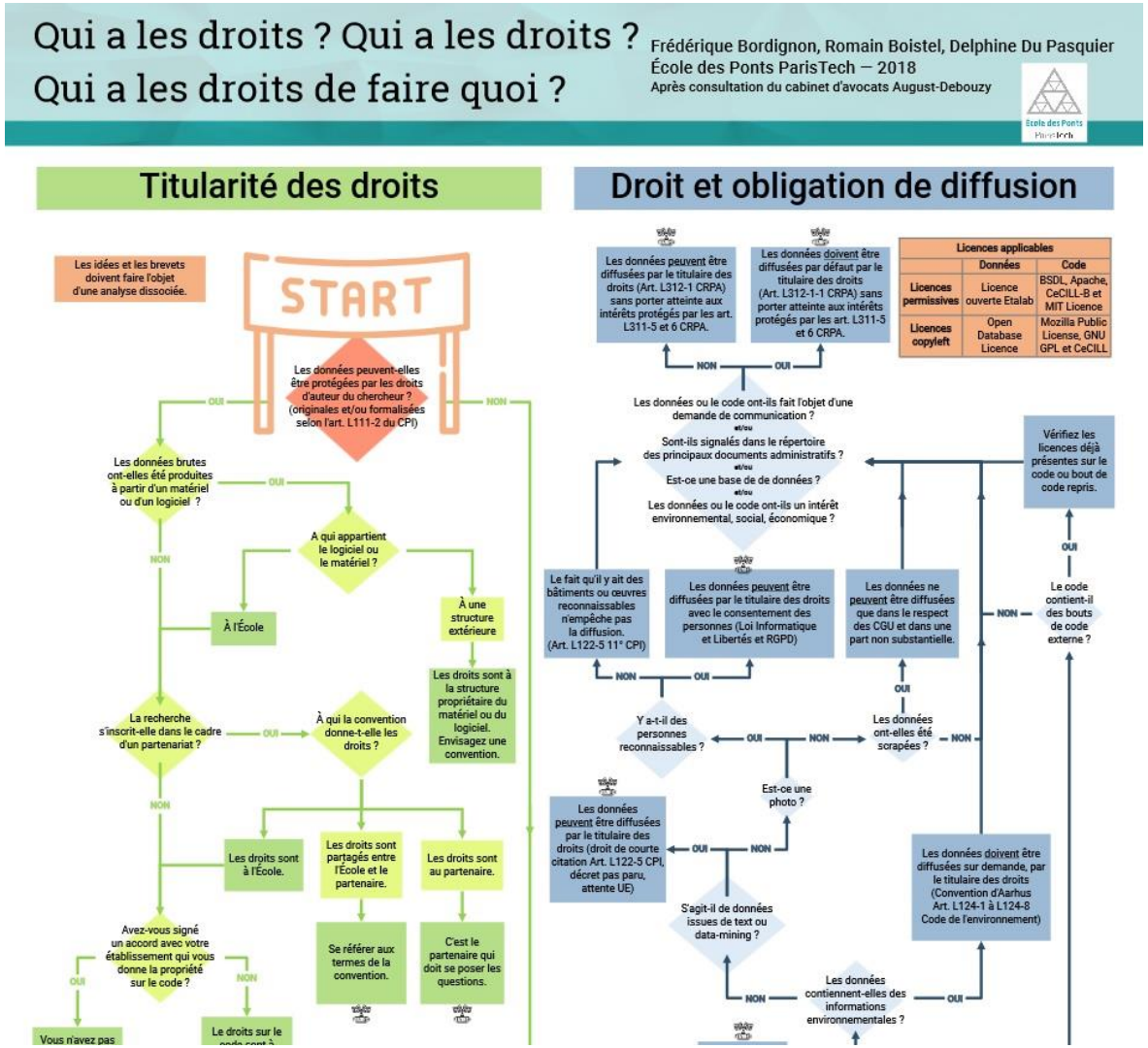
nakala

DRYAD

The Dataverse® Project

HAL archives-ouvertes.fr

Droit des données de la recherche



Travaux menés par l'Ecole des Ponts ParisTech.
(Frédérique Bordignon, Romain Boistel et Delphine Du Pasquier)

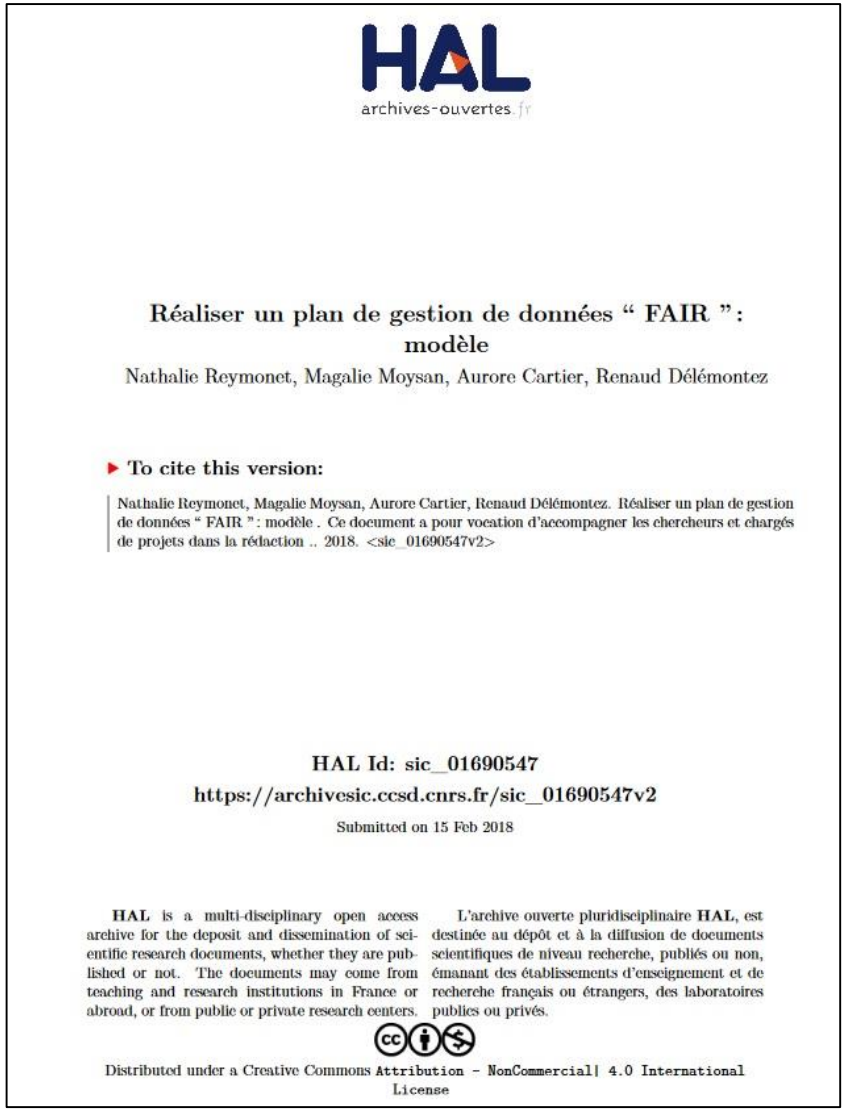
https://espacechercheurs.enpc.fr/sites/default/files/logigramme_a_plat.pdf

https://espacechercheurs.enpc.fr/fr/logigramme_dynamique

https://espacechercheurs.enpc.fr/sites/default/files/Synth%C3%A8se%20Data%20Questions%20juridiques%20Pole-IST.pdf

https://espacechercheurs.enpc.fr/fr/donnees-recherche-contexte-juridique

Guide de rédaction



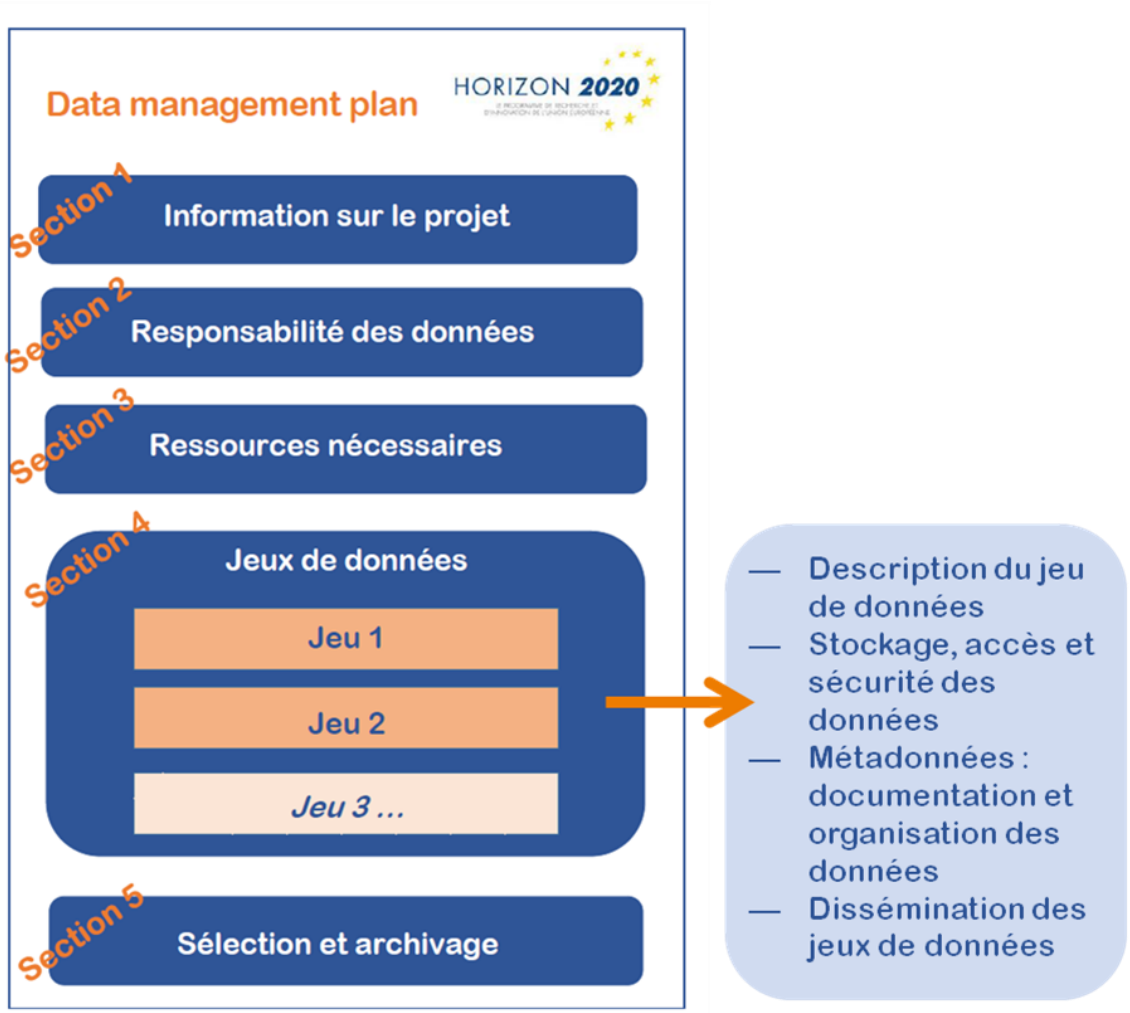
HAL
archives-ouvertes.fr

Réaliser un plan de gestion de données " FAIR ": modèle

Nathalie Reymonet, Magalie Moysan, Aurore Cartier, Renaud Délémontez

https://archivesic.ccsd.cnrs.fr/sic_01690547/document

Exemple de template



Section 1: informer sur le contexte administratif et scientifique du projet de recherche auquel est lié le DMP
Section 2: identifier la ou les personne(s) qui seront en charge l'application et de la mise à jour du DMP tout au long du projet
Section 3: estimer les compétences, ressources et coûts nécessaires à la mise en œuvre du DMP : gestion, curation, formation et archivage
Section 4: présenter le type de données du jeu qui seront produites et reçues dans le cadre du projet
Section 5: sélectionner et prévoir l'archivage à long terme des données ayant vocation à être conservées

Les sections 1,2,3 et 5 sont valables pour l'ensemble du projet.
La section 4 est propre à chaque jeu de données.

Enquête

Enquête menée par la Research Data Alliance, du 9 janvier au 25 mars 2019.
Objectif: comprendre les attentes sur le partage et la diffusion des plans de gestion de données

https://rd-alliance.org/group/dmp-common-standards-wg-exposing-data-management-plans-wg-active-data-management-plans-ig