Ecole Thématique DATA SDUE

Guide de Survie dans la jungle des données

en Sciences de l'Univers et de l'Environnement (SDUE) :

Comment gérer les données pour les valoriser?

Session «Choix de l'entrepôt de données pour dépôt et diffusion des données FAIR »

Françoise Genova

Choix d’un entrepôt

Françoise Genova

Un entrepôt … c'est quoi ?

Structure **spécifique à un domaine scientifique** pouvant contenir une ou plusieurs bases de données et regroupant un ou plusieurs jeux de données. L'entrepôt **les met à disposition pour leur utilisation.** Il les **organise** de manière logique avec les **métadonnées associées et en assure la conservation à long terme** (intégrité physique, archivage, *curation, preservation)*. Les entrepôts de données peuvent avoir des exigences spécifiques et/ou des restrictions statutaires concernant :

- le sujet ou le domaine de recherche ;
- la qualité des données ;
- l'origine des données ;
- la réutilisation et l'accès aux données ;
- les formats de fichier et la structure des données ;
- les types de métadonnées;

Un entrepôt de données doit à terme engager un processus de certification (type *CoreTrustSeal* : des données FAIR dans un entrepôt TRUST).

Définition Glossaire INSU

Choisir un entrepôt pour déposer des données

- Pas de réponse simple

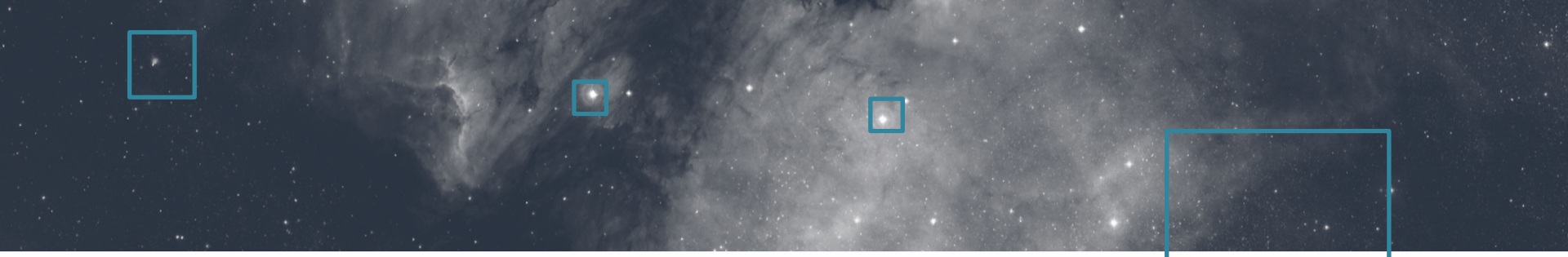
- Dans l'idéal, forte composante thématique
  - Un entrepôt connu dans la communauté
  - Un entrepôt qui a la confiance des utilisateurs

Confiance/Trust

- La confiance est un composant essentiel de la science ouverte
  - Pour les personnes qui partagent leurs données
  - Pour les personnes qui utilisent des donnes
  - Pour les financeurs des éléments qui rendent la science ouverte possible
- Comment définir un « entrepôt de données confiance »/ « trustworthy data repository »?

La certification des entrepôts de données

- Définir les critères pertinents pour déterminer si un entrepôt est de confiance
- Pour les entrepôts mis en œuvre par des structures de recherche: Certification CoreTrustSeal



LA CERTIFICATION DES ENTREPOTS DE DONNÉES: UN PEU D'HISTOIRE

Merci à Ingrid Dillo et Hervé L'Hours

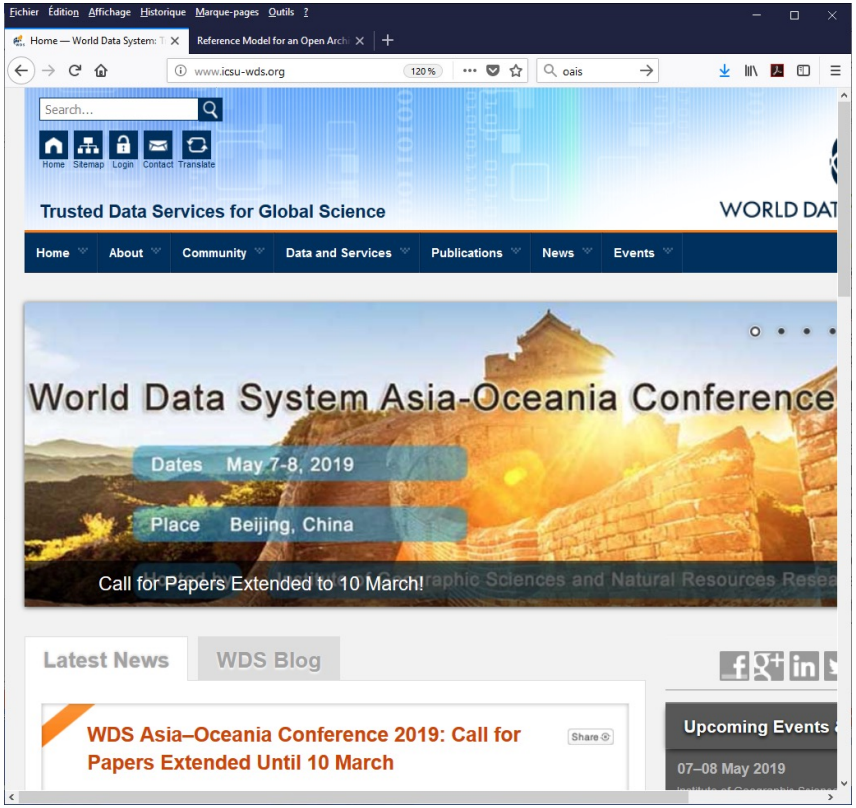
The certification landscape EN 2015

- 4 certification standards available

DIN 31644

ISO 16363

Le World Data System (WDS)

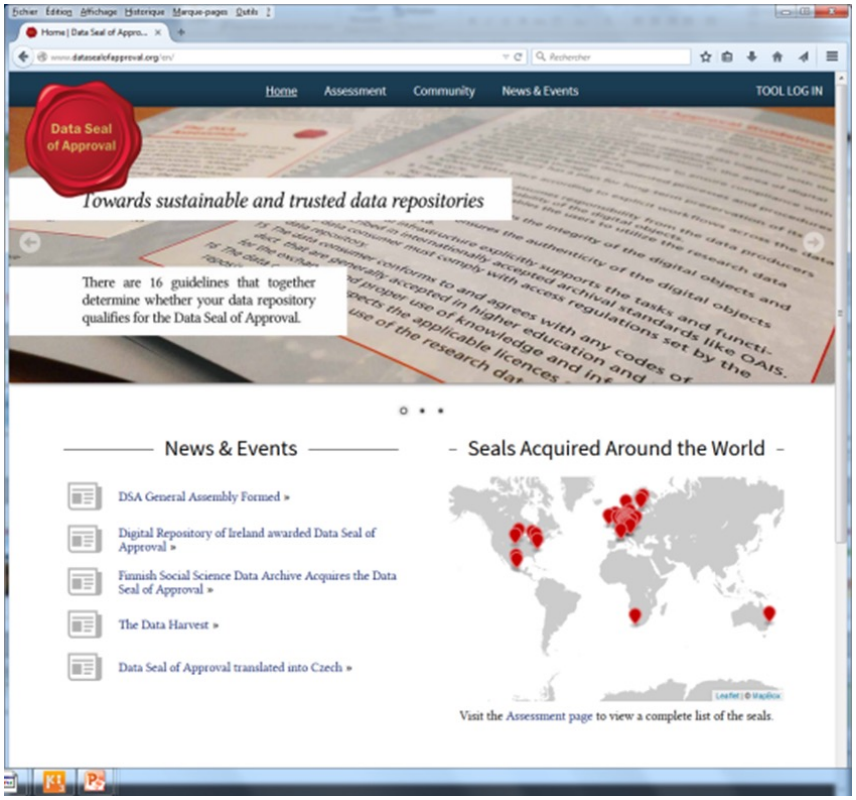


- Créé en 2008 par l'ICSU
- Au départ essentiellement données sur la planète (et astronomie) mais ouvert à tous
- *Promoting universal and equitable access to, and long-term stewardship of, quality-assured scientific data and data services, products, and information covering a broad range of disciplines from the natural and social sciences, and humanities.*
- *Coordinates trusted scientific data services for the provision, use, and preservation of relevant datasets*

European certification framework (pre-RDA)

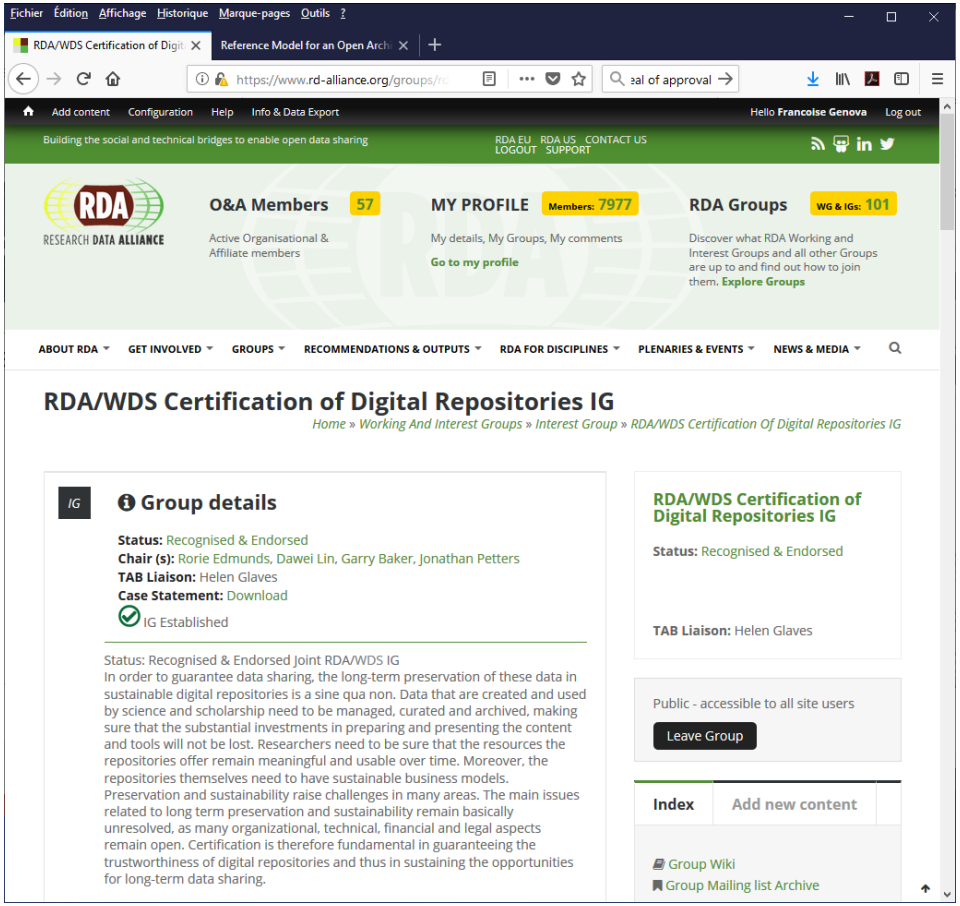
- **Basic Certification** is granted to repositories which obtain DSA certification
- **Extended Certification** is granted to Basic Certification repositories which *in addition* perform a structured, externally reviewed and publicly available self-audit based on DIN 31644/nestorSeal
- **Formal Certification** is granted to repositories which *in addition t*o Basic Certification obtain full external audit and certification based on ISO 16363

Le Data Seal of Approval

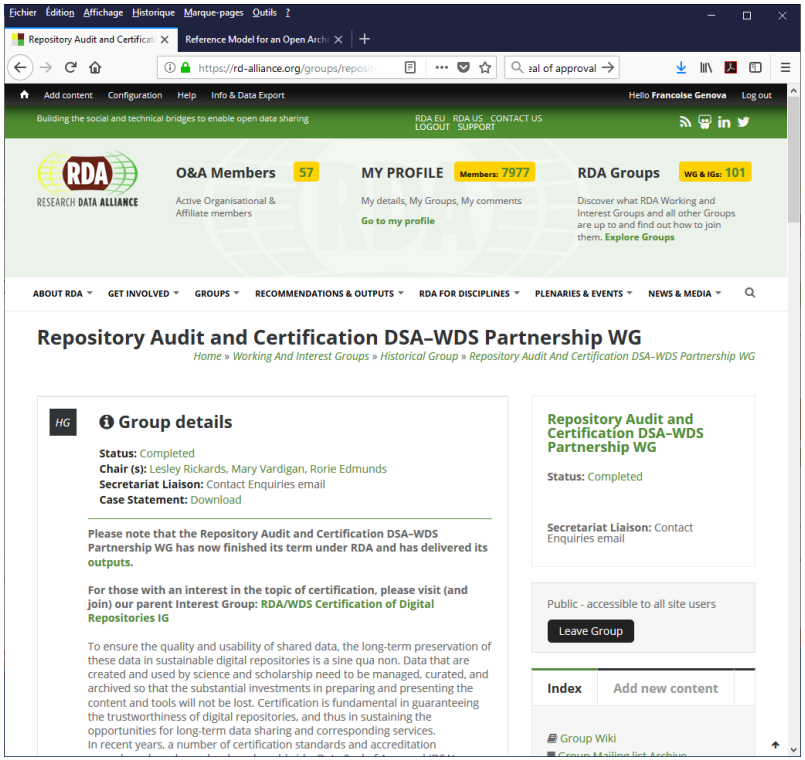


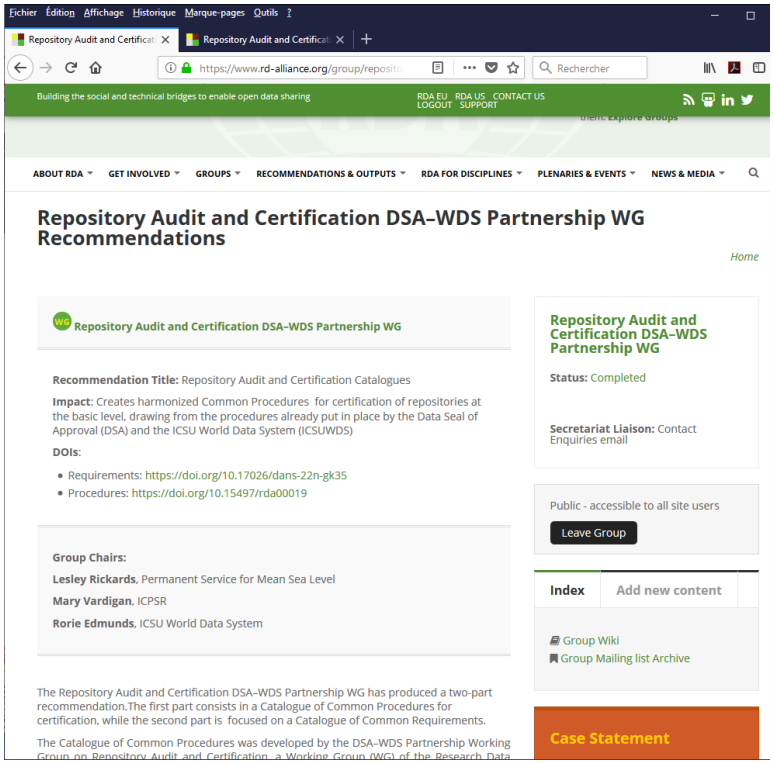
- Plutôt Humanités
- Plus entrepôts de données que services
- Le CDS a été le premier centre certifié DSA du domaine des sciences physiques (en 2014)

Dans la RDA, dès 2013

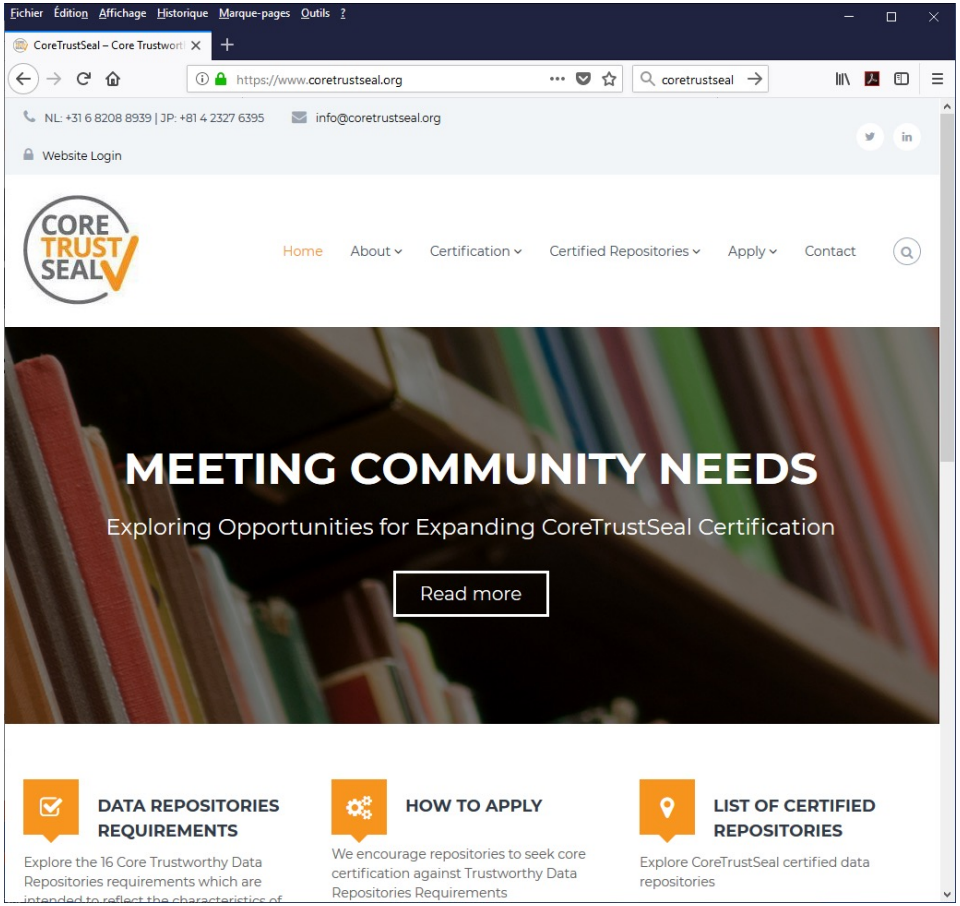


RDA: DSA + WDS (2016)





DSA + WDS => CoreTrustSeal (CTS)



Le modèle de certification européen

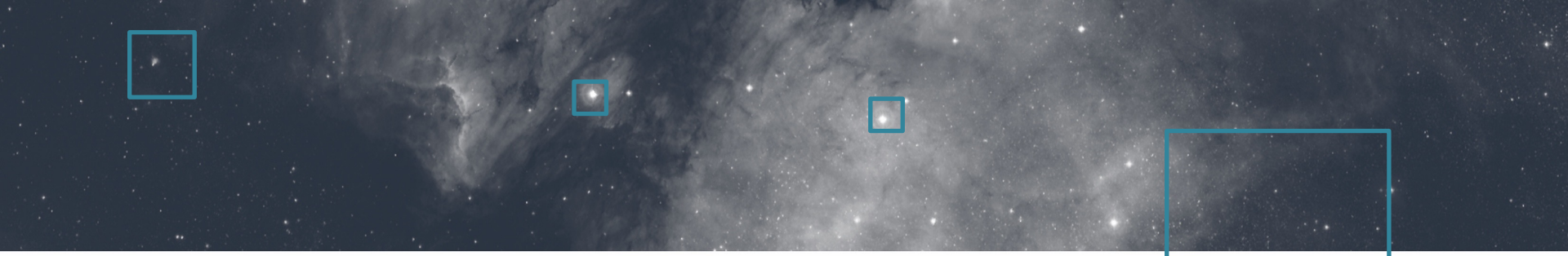
Certification formelle ISO 16363:2012
(2 entrepôts ont été certifiés – 100+ critères)
https://www.iso.org/standard/56510.html/

Certification étendue DIN 31644
(4 entrepôts certifiés, 34 critères)
www.langzeitarchivierung.de

Certification CoreTrustSeal
(132 entrepôts, 16 critères)
www.CoreTrustSeal.org

Centres de données de confiance - Trustworthy Data Repositories

Thanks to Mustapha Mokrane - NestorSeal and ISO numbers updated 22 January 2021, CTS 11 October 2022



LES CRITÈRES DE LA CERTIFICATION CORETRUSTSEAL

Les critères de certification

- Informations de base et contexte
- Infrastructure organisationnelle
- Gestion des objets numériques
  - Le modèle OAIS est sous-jacent
- Technologies de l'information et sécurité

Information de base et contexte

- re3data identifier
- Repository type
  - Generalist repository
  - Specialist repository
- Overview
- Designated community
- Levels of curation

  A. Content distributed as deposited

  B. Basic curation – e.g. brief checking, addition of basic metadata or documentation

  C. Enhanced curation – e.g. conversion to nex formats during ingest, enhancement of documentation and metadata

  D. Data-level curation- as in C above, but with additional editing of deposited data
- Cooperation and outsourcing to third parties, partners and host organisations
- Applicants renewing their CoreTrustSeal certification: Summary of significant changes since last application

Infrastructure organisationnelle

R01 Mission & Scope

The repository has an explicit mission to provide access to and preserve digital objects

R02 Rights management

The repository maintains all applicable rights and monitor compliance

R03 Continuity of service

The repository has a plan to ensure ongoing access to and preservation of its data and metadata

R04 Legal & Ethical

The repository ensures to the extent possible that data are created, preserved, accessed and used in compliance with legal and ethical norms

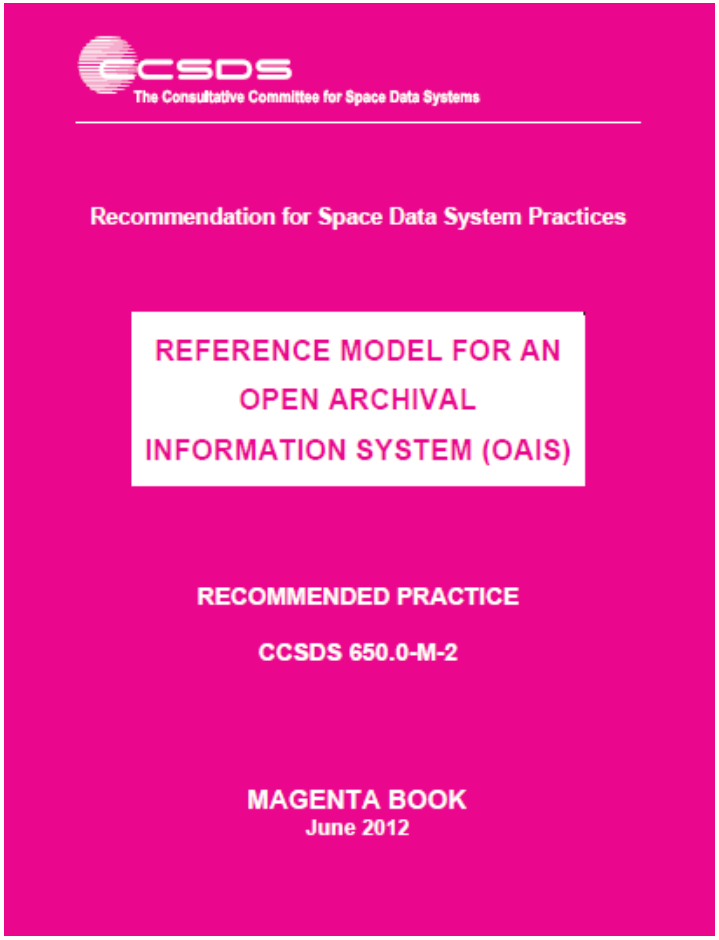
R05 Governance & Resources

The repository has adequate funding and sufficient numbers of staff managed through a clear system of governance to effeectively carry out the mission

R06 Expertise & Guidance

The repository adopts mechanisms to secure ongoing expertise, guidance and feedback – either in-house, or external

Le modèle OAIS



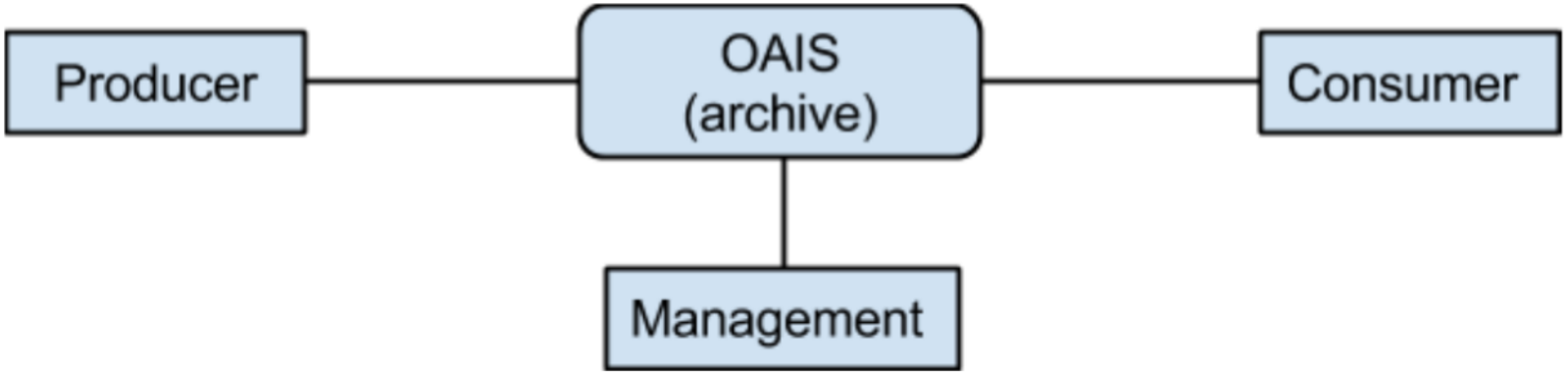
- OAIS – Open Archive Information System

https://public.ccsds.org/Pubs/650x0m2.pdf

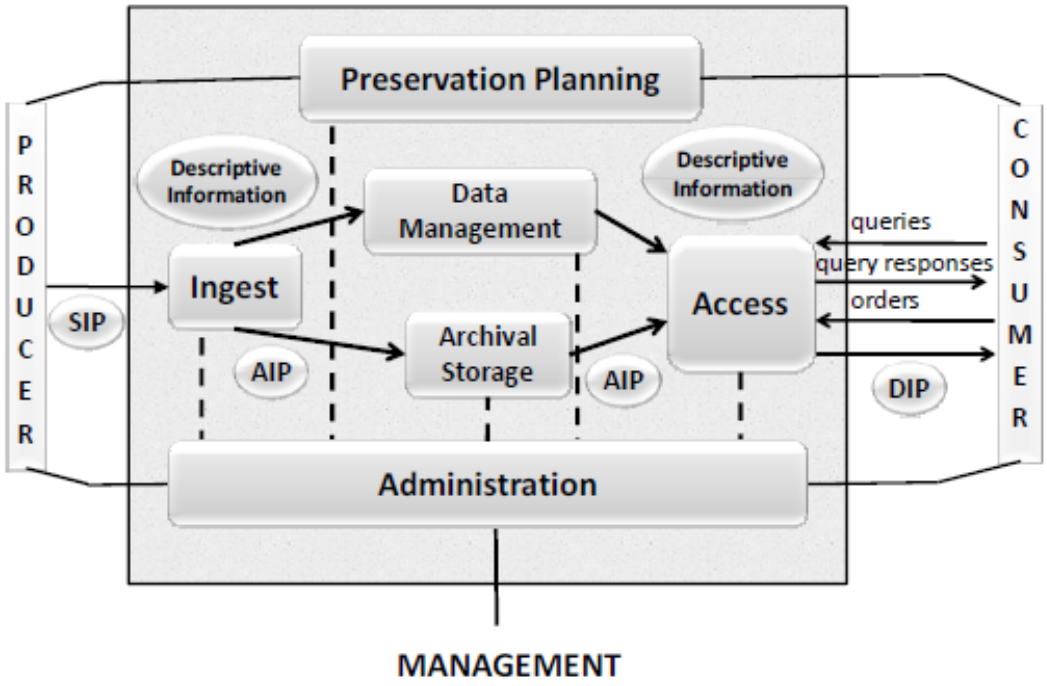
- Site en français

https://www.cines.fr/archivage/un-concept-des-problematiques/le-modele-de-reference-loais/

L'environnement d'une archive OAIS

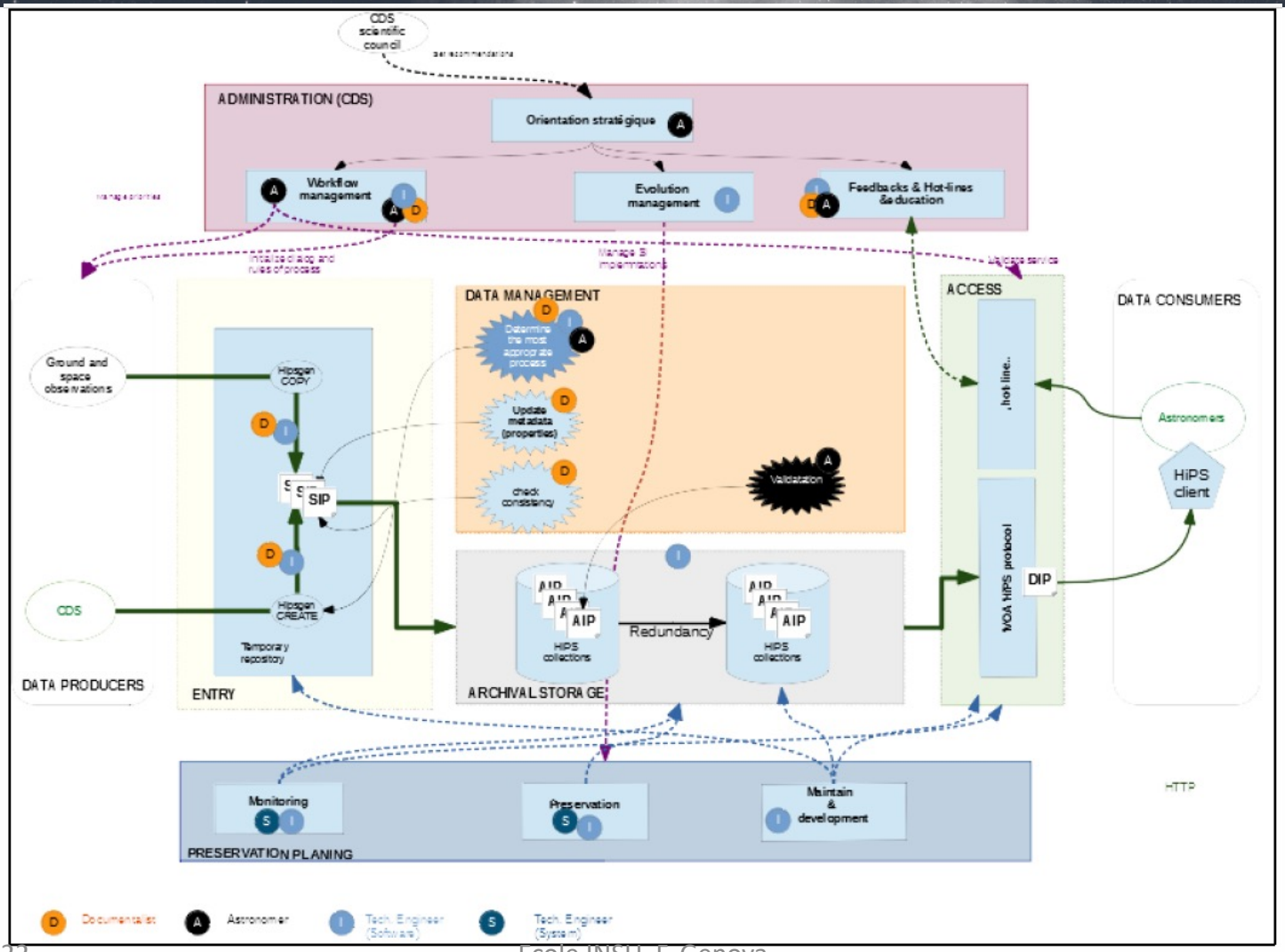


Les entités fonctionnelles de l'OAIS

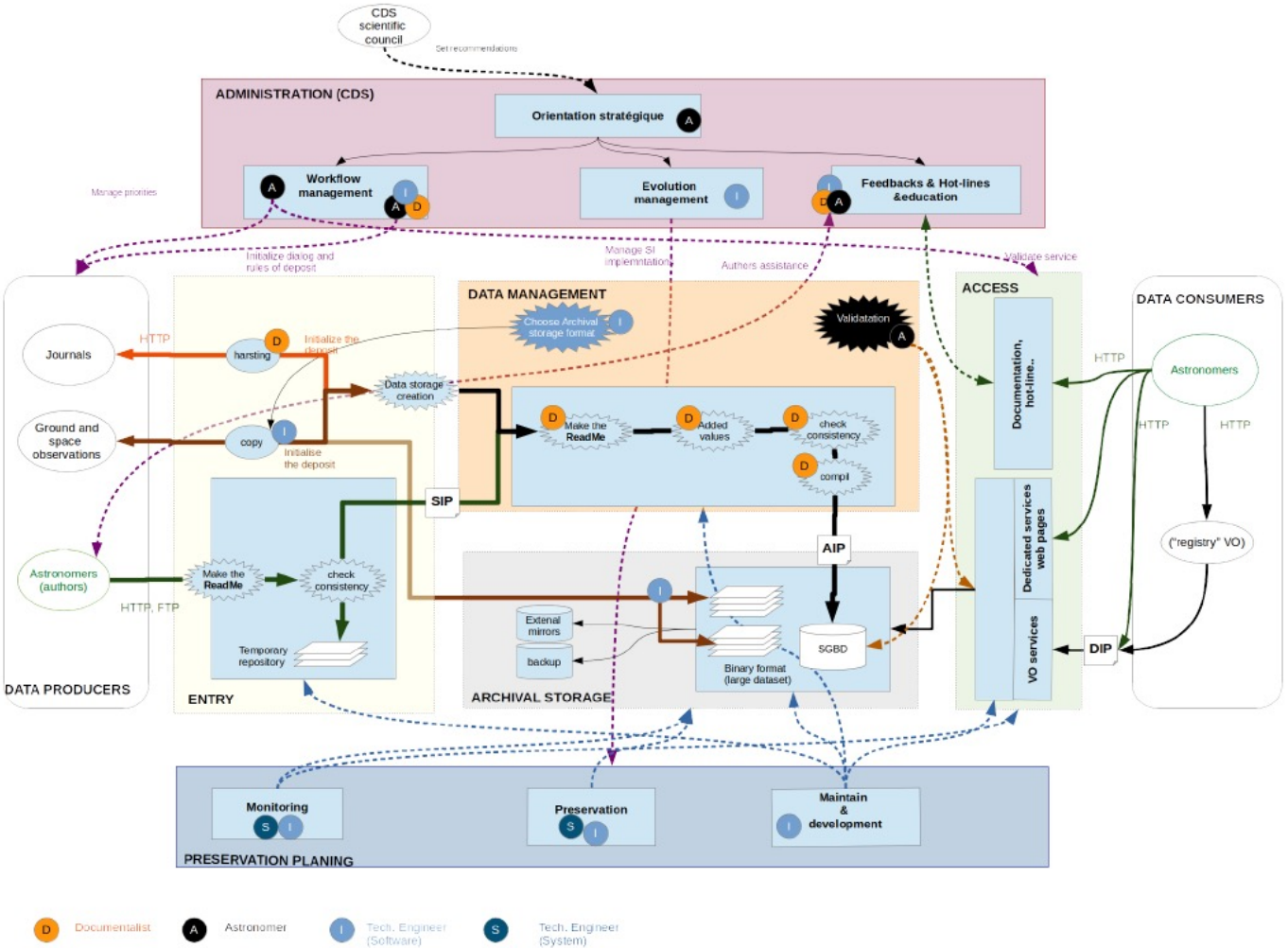


SIP: Submission Information Package
AIP: Archival Information Package
DIP: Dissemination Information Package

L'entrepôt Aladin du CDS



L'entrepôt VizieR du CDS



Gestion des objets numériques

R07 Provenance and authenticity

The repository guarantees the autheticity of the digital objects and provides provenance information

R08 Deposit & Appraisal

The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for users

R09 Preservation plan

The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way

R10 Quality Assurance

The repository addresses technical quality and standards compliance, and ensures that sufficient information is available for end users to make quality-related evaluations

R11 Workflows

Digital object management takes place according to defined workflows from deposit to access

R12 Discovery and Identification

The repository enables users to discover the digital objects and refer to them in a persistent way through proper citation

R13 Reuse

The repository enables reuse of the digital objects over time, ensuring that appropriate information is available to support understanding and use

Technologies de l'information et sécurité

R14 Storage & Integrity

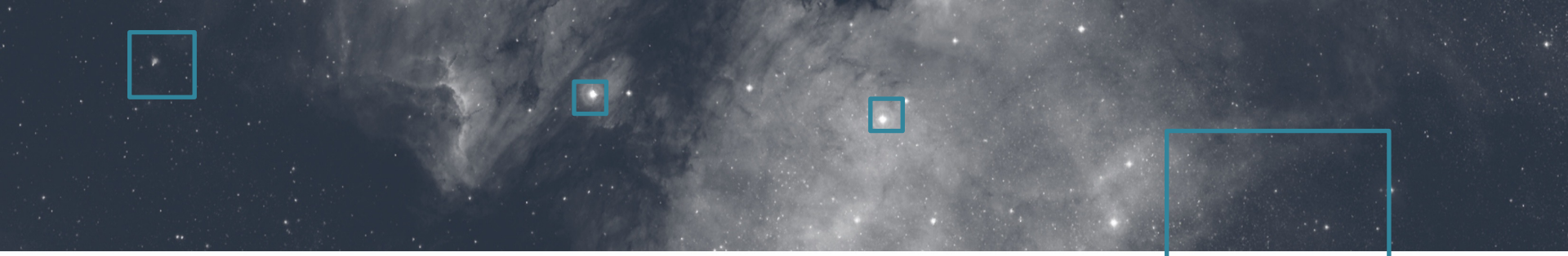
The repository applies documented processes to ensure data and metadata storage and integrity

R15 Technical infrastructure

The repository is managed on well-supported operating systems and other core infrastructural software and hardware appropriate to the services it provides to its Designated Community

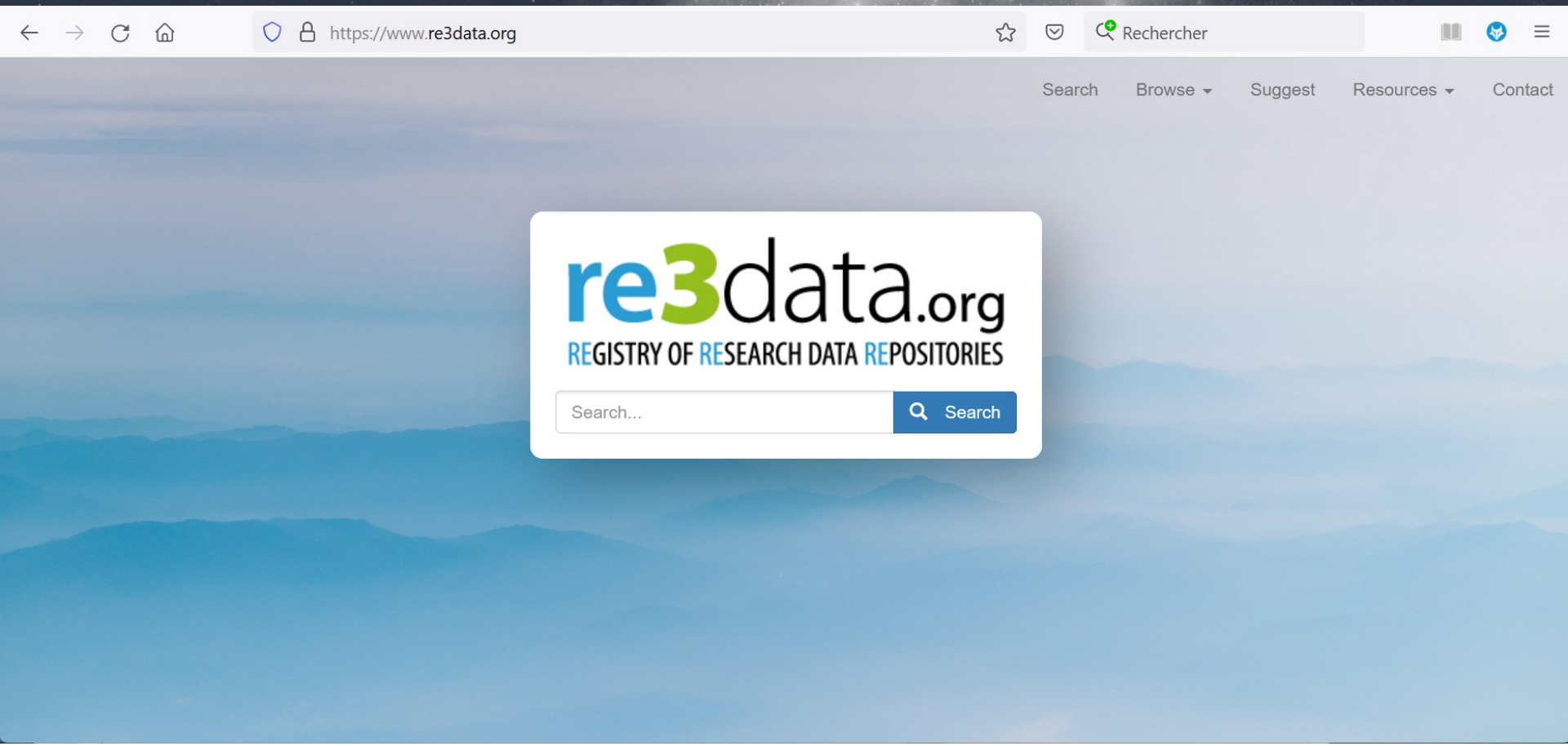
R16 Security

The repository protects the facility and its data, metadata, products, services, and users



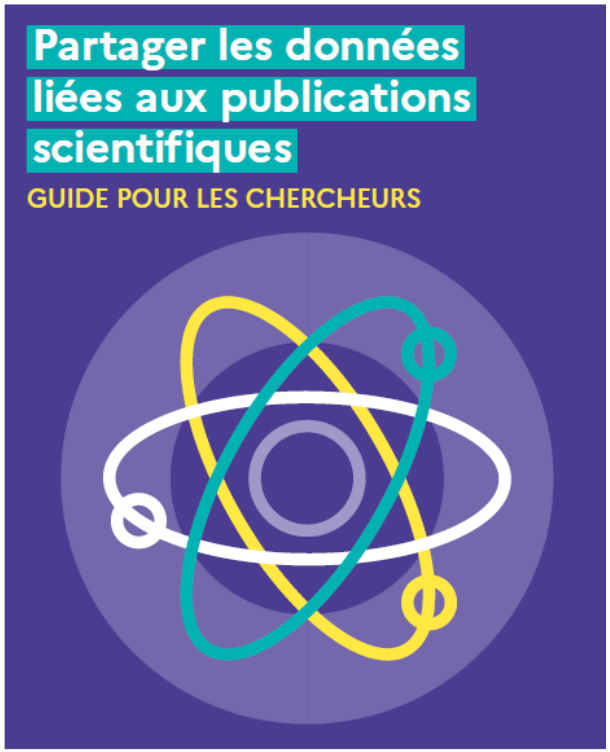
ET QUAND IL N'Y A PAS D'ENTREPÔT CERTIFIÉ ADÉQUAT?

re3data



https://www.re3data.org/

Guide publié par le CoSO



Dans le guide du CoSO

Il est recommandé **de ne pas confier les données à partager aux éditeurs des revues** qui proposent de les publier sous forme de « *supplementary data* » ou de « *supplementary materials* » associés à l'article. Une telle publication se fait encore souvent dans un format et un environnement qui ne permettent pas de documenter correctement les données et rendent difficile leur réutilisation. Elle peut aussi s'accompagner d'une demande de transfert exclusif de droits contraire à la loi française et à l'esprit de la science ouverte. Enfin, dans certains cas, elle contribue à rendre les utilisateurs captifs au sein d'environnements maîtrisés par de grands acteurs commerciaux de l'édition scientifique.

Il est donc plutôt recommandé d'utiliser pour le partage des données des **entrepôts de données institutionnels, généralistes ou disciplinaires**, qui permettent d'éviter ces écueils et offrent un environnement dédié à la documentation, l'ouverture et la réutilisation de la donnée de recherche. Établir correctement le **lien entre le jeu de données déposé dans l'entrepôt et l'article disponible sur une plateforme de publication** devient alors une nécessité et une démarche à anticiper.

Choix de l'entrepôt

- Dans le cas de disciplines structurées pour le partage des données (astronomie, génomique, etc…), les producteurs de données ont à disposition des **entrepôts spécifiques à leur discipline**. Ils utiliseront alors naturellement l'ensemble des standards et bonnes pratiques déjà en place pour documenter et mettre en forme leurs données. La pratique de sa communauté est le meilleur guide mais des annuaires de ces entrepôts existent[4].

=> Discuter avec les chercheurs

- En alternative, les producteurs de données pourront se tourner vers l'**entrepôt institutionnel** auquel ils sont affiliés, s'il existe, ou utiliser l'**entrepôt pluridisciplinaire Recherche Data Gouv**. Dans ces deux cas, des exigences minimales seront imposées par les entrepôts et la charge de s'assurer de la qualité de la documentation des données sera davantage portée par le déposant.

L'entrepôt national Recherche Data Gouv :

La plateforme nationale Recherche Data Gouv propose un entrepôt de données pluridisciplinaire qui sera opérationnel dès 2022 : il assure la souveraineté française sur les données, est conforme aux droits français et communautaire, garantit la pérennité et l'indexation des données stockées, suivant les principes FAIR. C'est l'entrepôt de choix quand aucun entrepôt disciplinaire n'existe.

Quel que soit l'entrepôt choisi pour partager les données, celui-ci se doit en particulier d'offrir les fonctionnalités suivantes :

- L'assignation d'un **identifiant pérenne** (*Persistent Identifier* : PID) de type DOI qui permet de citer les données (par exemple http://dx.doi.org/10.15497/RDA00027) et constitue la brique de base pour établir le lien avec d'autres produits de la recherche comme les publications.
- La **description des données** à un niveau suffisant pour en faciliter la découverte, la compréhension et la réutilisation (métadonnées descriptives standardisées, vocabulaires disciplinaires contrôlés).
- L'utilisation de **licences** et la définition de **règles d'accès** permettant d'inscrire la réutilisation dans un cadre légal bien défini et compatible avec le droit français et européen[5].
- Une **durée de conservation** minimale de plusieurs années, cohérente avec la politique de conservation des données de l'établissement de rattachement.

Comment choisir l'entrepôt où déposer?

Entrepôt thématique

Entrepôt certifié

Entrepôt à portée nationale

Entrepôt hébergé par un DataCenter national labellisé

Entrepôt généraliste

Entrepôt non certifié

Entrepôt très local (OSU, labo,..)

Entrepôt hébergé par un serveur tour dans le local informatique au sous-sol du bâtiment hébergeant l'OSU X

Liens utiles

- Recommandation RDA sur la certification
  https://www.rd-alliance.org/group/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg/outcomes/dsa-wds-partnership
- CoreTrustSeal
  Site web https://www.coretrustseal.org/
  Critères 2023-2025 https://www.coretrustseal.org/why-certification/requirements/
- re3data
  https://www.re3data.org/
- Guide CoSO Partager les données
  https://www.ouvrirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guide-pour-les-chercheurs/

Un entrepôt / Data Repository ce N'est PAS :

- **un jeu de données / *Dataset*** = *Toute collection organisée de données dans un format numérique, définie par un thème ou une catégorie*
- **une base de données / *Database*** = *collection de jeux de données et d'informations organisées afin d'être facilement consultables (searchable), gérables et mises à jour.*
- **un *Data Warehouse*** = *base de données relationnelle hébergée sur un serveur dans un DataCenter ou un Cloud.*
- **un *Data Lake*** = *désigne un espace de stockage global des informations présentes au sein d'une organisation (plat comme un lac, pas de priorité)*
- **un centre de données / *DataCenter*** = *désigne un lieu physique environné pour l'hébergement de différents équipements informatiques*
- un site Web en ligne (html clickable, FTP, etc)