La certification des entrepôts de données

Françoise Genova, CDS/Observatoire Astronomique de Strasbourg, RDA France

Aude Chambodut, EOST, CNRS INSU, Board CTS

Olivier Rouchon, CINES, FAIRsFAIR, Board CTS

La certification des entrepôts de données

- Pourquoi?
- Les cadres de certification ‘de base’
- Exemple d’auto-évaluation
- Les critères du CoreTrustSeal
- Conclusions

- *Tout au long de l’exposé, l’exemple du Centre de Données astronomiques de Strasbourg (CDS)*

La certification, pourquoi?

RESEARCH DATA ALLIANCE

Certification

from I. Dillo and H. L'Hours

research data sharing without barriers

rd-alliance.org


Perhaps the biggest challenge in sharing data is trust: how do you create a system robust enough for scientists to trust that, if they share, their data won't be lost, garbled, stolen or misused?
The Data Harvest:
How sharing research data can yield knowledge, jobs and growth
An RDA Europe Report
December 2014

Trust

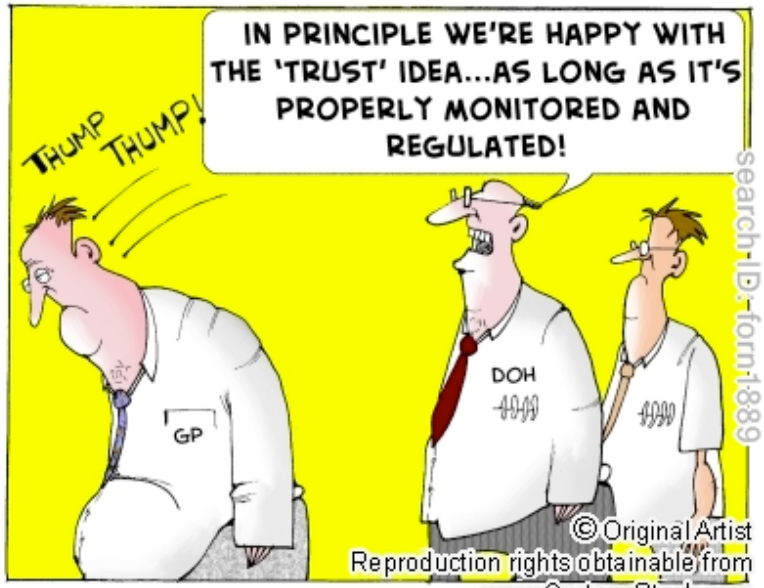
Trust is at the very heart of storing and sharing data

- Users
- Depositors
- Funders

What is trust built on?

- Dedicate yourself (mission statement)
- Do what you promise (stable, sincere and competent reputation)
- Be transparent (peer review, get certified)

"They don't trust each other to share research."



RDA
RESEARCH DATA ALLIANCE

What is a trustworthy repository?

- mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future
- constant monitoring, planning, and maintenance
- understand threats to and risks within its systems
- **regular cycle of audit and/or certification**

What is a trustworthy repository?

- mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future
- constant monitoring, planning, and maintenance
- understand threats to and risks within its systems
- **regular cycle of audit and/or certification**

?

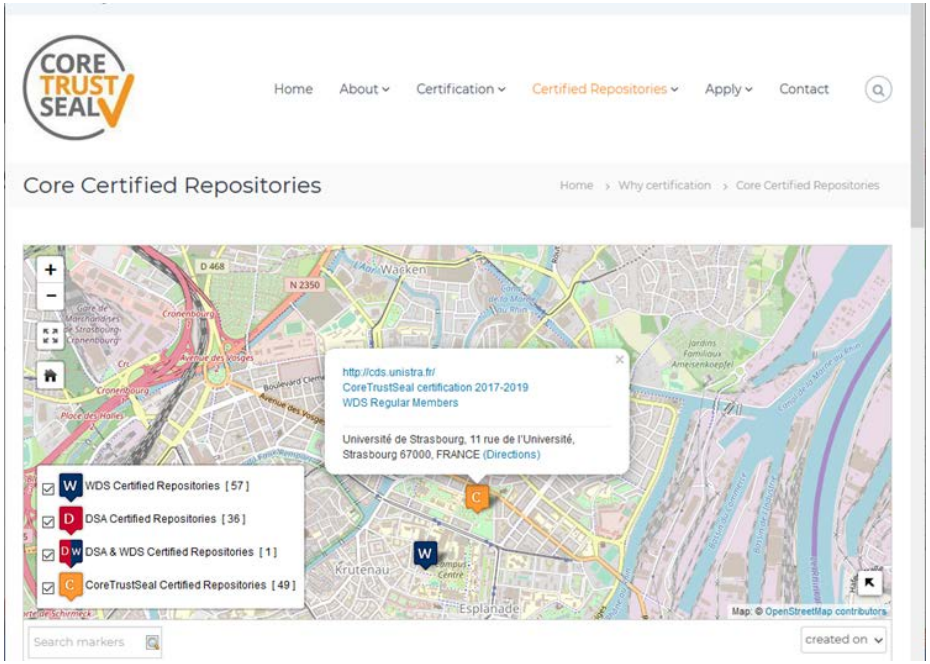
Pourquoi une certification formelle?

- Assurer que le centre est « de confiance »
- Mais… il a peut-être déjà la confiance de ses utilisateurs…
- L'exemple du CDS
  - Crée en 1972
  - Centre de données de référence pour la communauté astronomique internationale
  - Infrastructure de Recherche sur la Feuille de Route nationale
  - 1 000 000+ requêtes/jour sur les services

Oui, pourquoi?

- Critères établis par des personnes compétentes et applicables quel que soit le cadre disciplinaire
- Au préalable, auto-évaluation selon les critères, qui permet de vérifier l'organisation et les process et d'identifier des améliorations possibles
- Evaluation externe par des personnes compétentes
- Le dépôt dans un centre de données certifié est un point important dans les Data Management Plans (DMP)/Plans de Gestion des Données (PGD)

Le CDS a été/est …

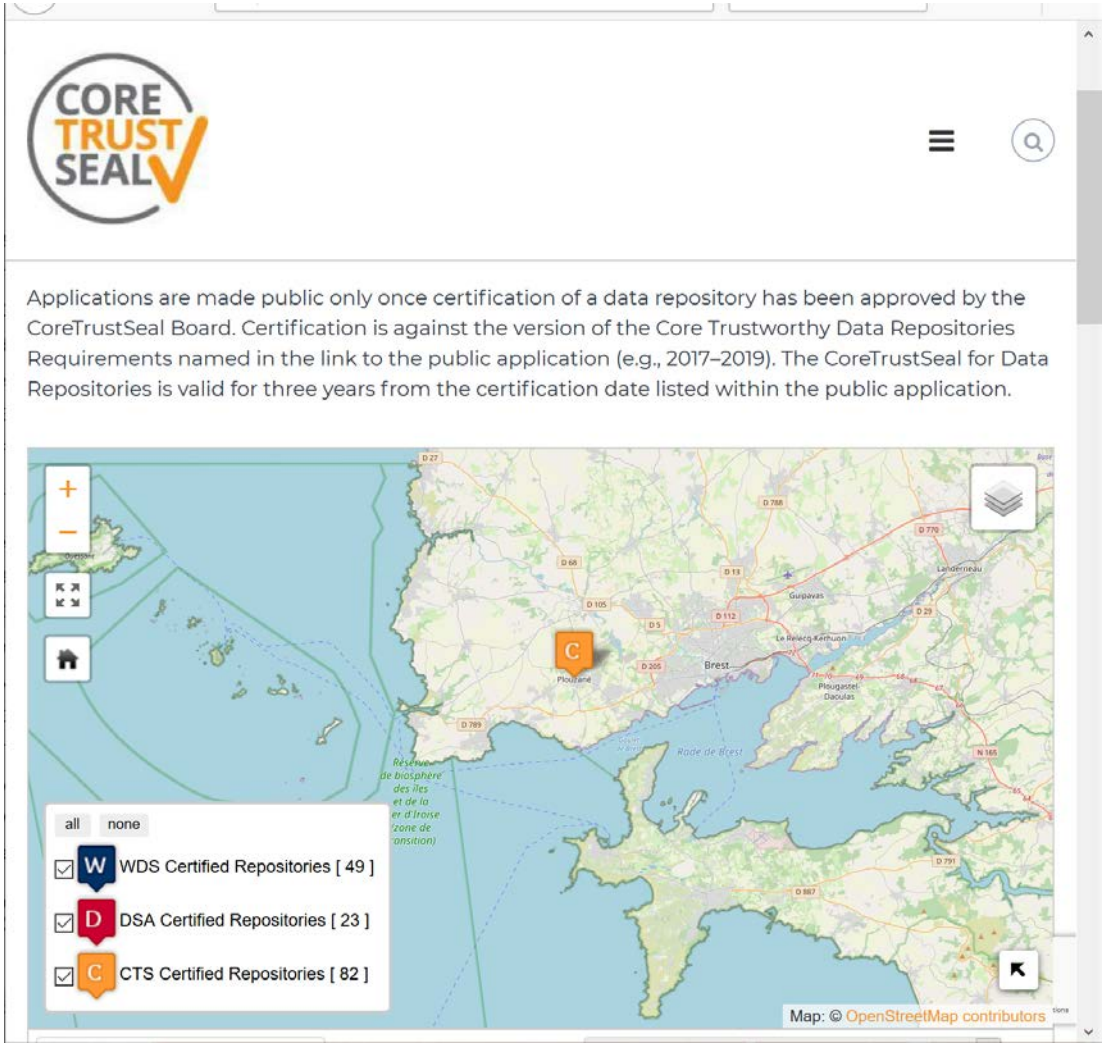


- *Regular Member* du *World Data System* - WDS
- Certifié par le *Data Seal of Approval* - DSA
- Certifié par *CoreTrustSeal* - CTS

- Document produit pour la certification CoreTrustSeal:

https://www.coretrustseal.org/wp-content/uploads/2019/02/Strasbourg-Astronomical-Data-Centre.pdf

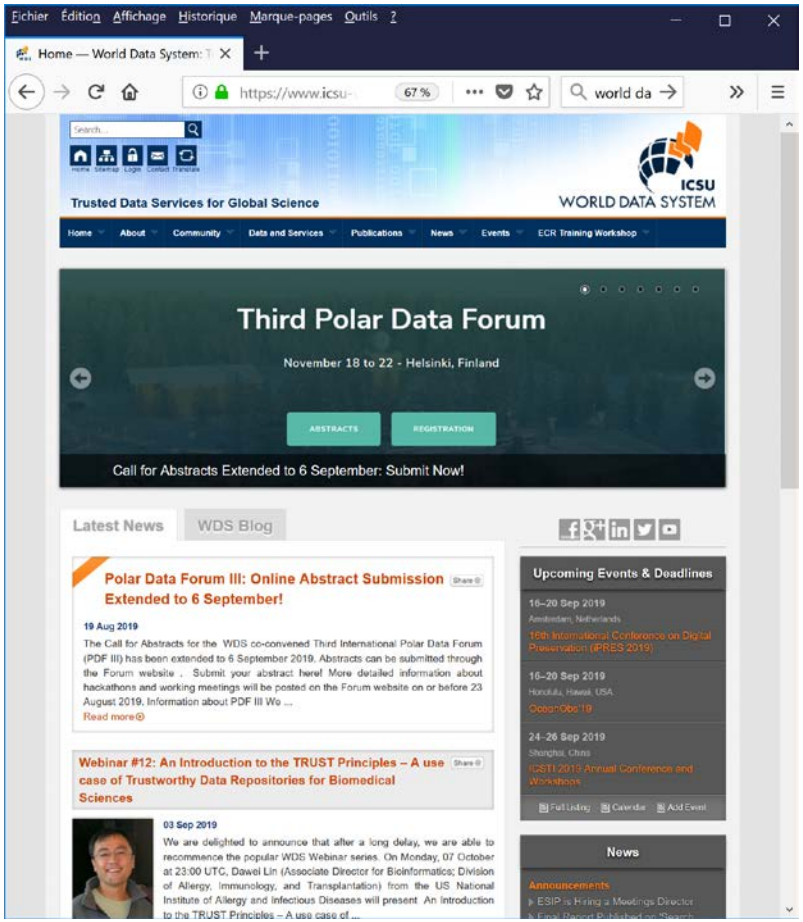
… ainsi que IFREMER-SISMER



Les cadres de certification ‘de base’

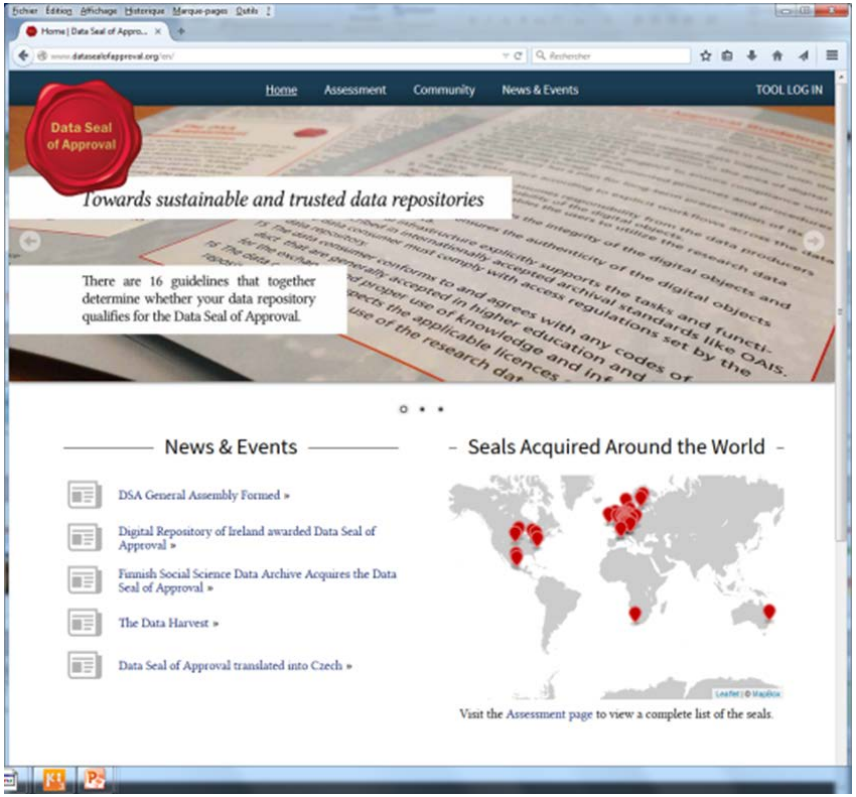
Un peu d’histoire…

Le World Data System (WDS)



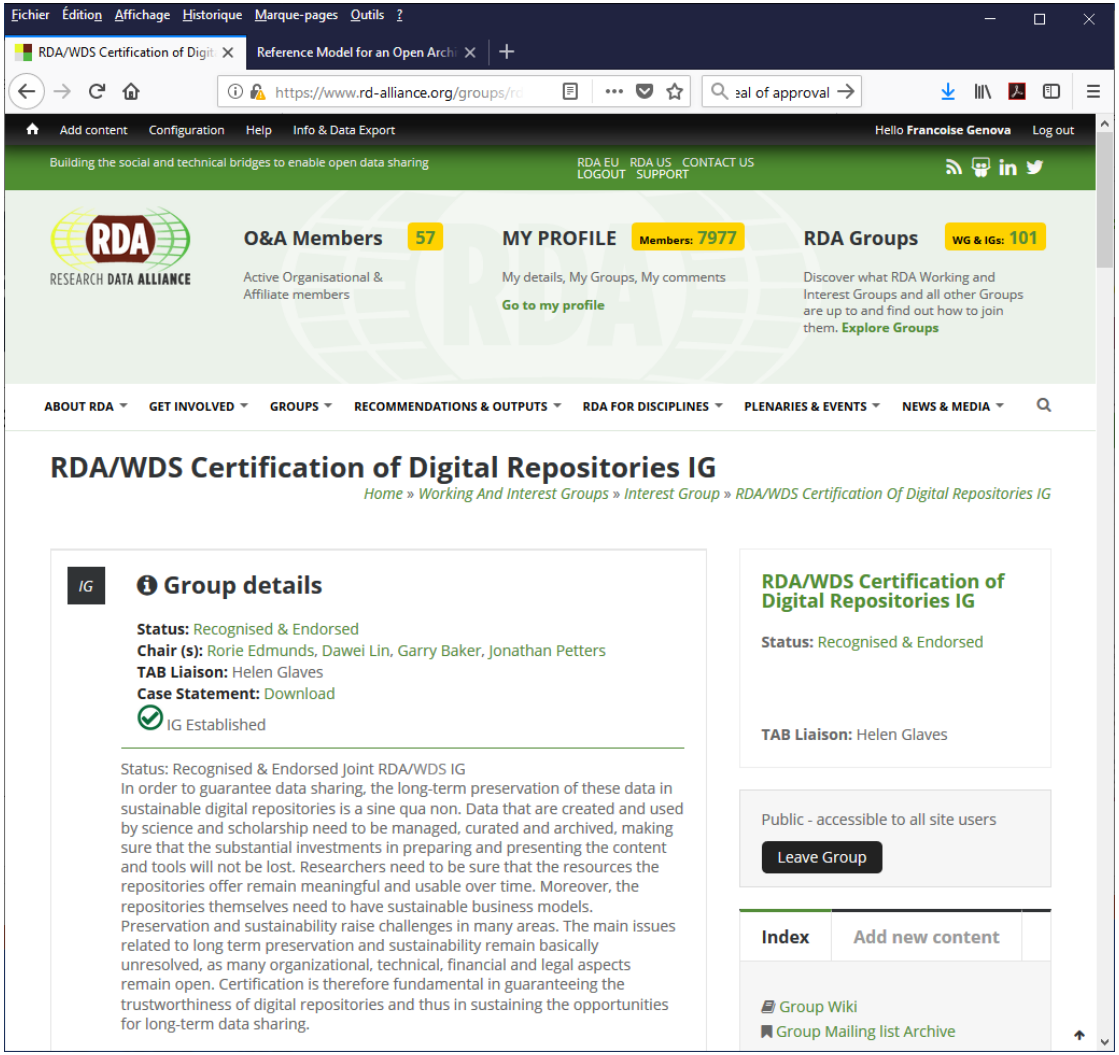
- Créé en 2008 par l'ICSU (=ISC)
- Essentiellement au départ données sur la planète (et astronomie) mais ouvert à tous
- *Promoting universal and equitable access to, and long-term stewardship of, quality-assured scientific data and data services, products, and information covering a broad range of disciplines from the natural and social sciences, and humanities.*
- *Cooordinates trusted scientific data services for the provision, use, and preservation of relevant datasets*
- CDS membre du WDS depuis 2012

Le Data Seal of Approval

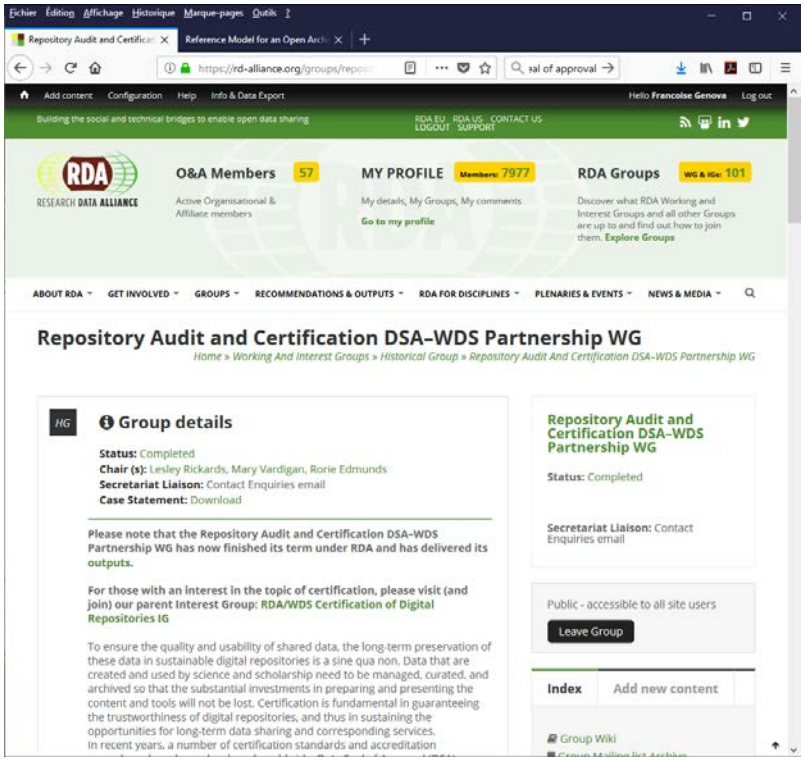


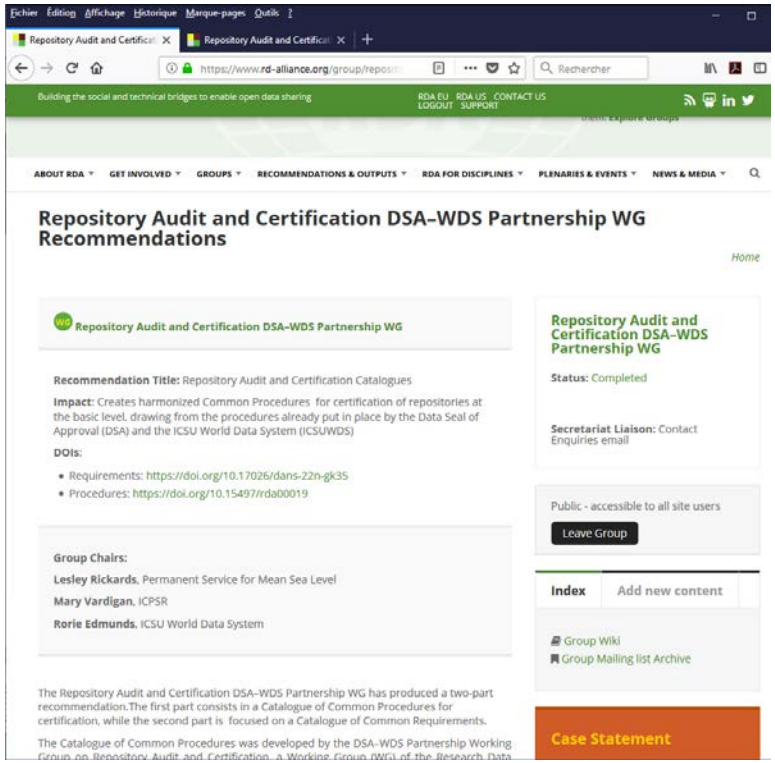
- Plutôt Humanités
- Plus dépôts de données que services
- Le CINES a été le premier centre français certifié DSA, le CDS a été le second en France et le premier centre certifié du domaine des sciences physiques (en 2014)

Dans la RDA, dès 2013

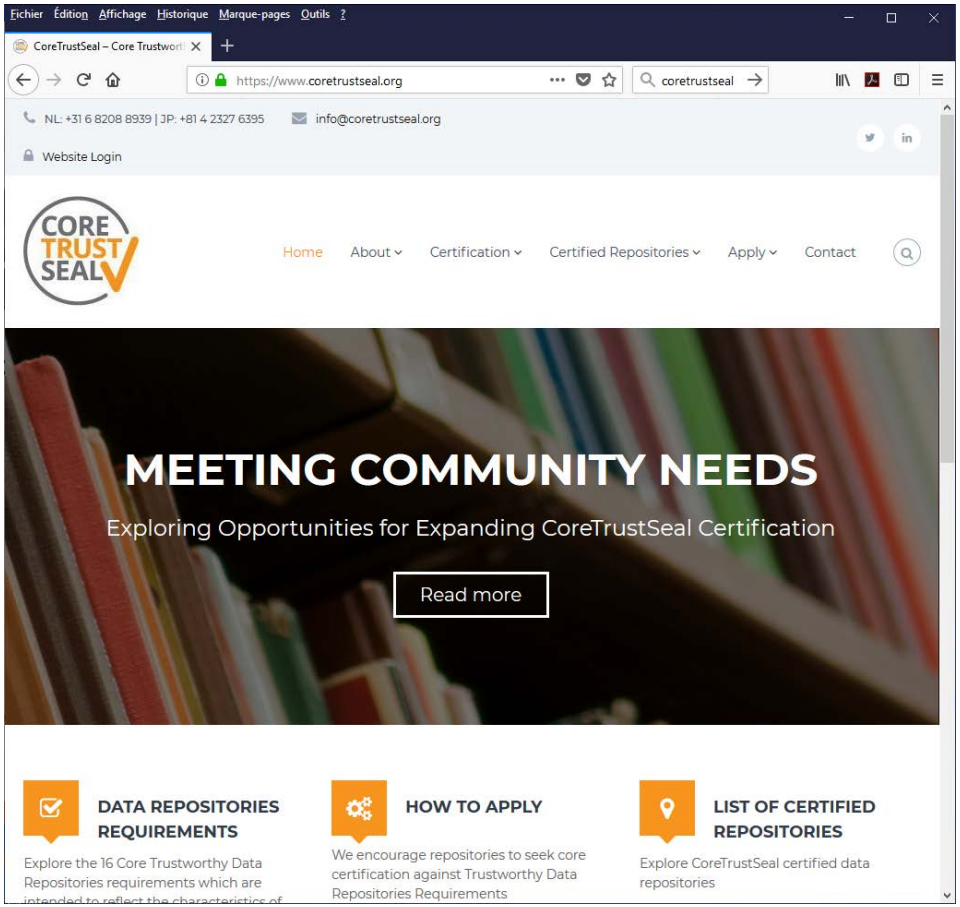


RDA: DSA + WDS (2016)





DSA + WDS = CoreTrustSeal (CTS), 2017



Le modèle de certification européen

Certification formelle ISO 16363:2012
(1 Centre certifié – US Gov Publishing Office, 100+ critères) https://www.iso.org/standard/56510.html/

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

Certification CoreTrustSeal
(170+ centres, 16 critères)
www.CoreTrustSeal.org

Centres de données de confiance - Trustworthy Data Repositories

Thanks to Mustapha Mokrane (DANS)- NestorSeal and ISO numbers updated 22 January 2021, CTS 12 October 2021

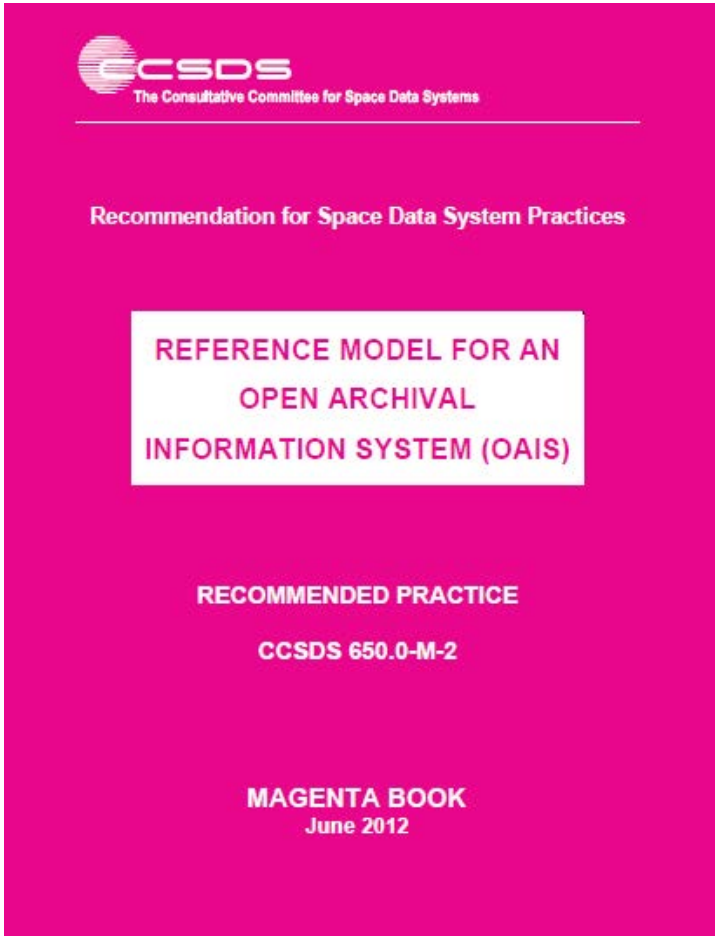
Exemple d'auto-évaluation

Le Centre de Données astronomiques de Strasbourg (CDS)

Auto-evaluation

- Questionnaire à remplir
- Il faut les compétences
  - De la direction (mission, organisation, …)
  - Des personnes en charge du contenu
  - Des personnes en charge de l'informatique
- Pour le CDS: un travail d'équipe qui a impliqué la direction, les documentalistes, l'informaticien en charge du service et l'ingénieur système

Description des process du CDS



- Basé sur le modèle OAIS – Open Archive Information System

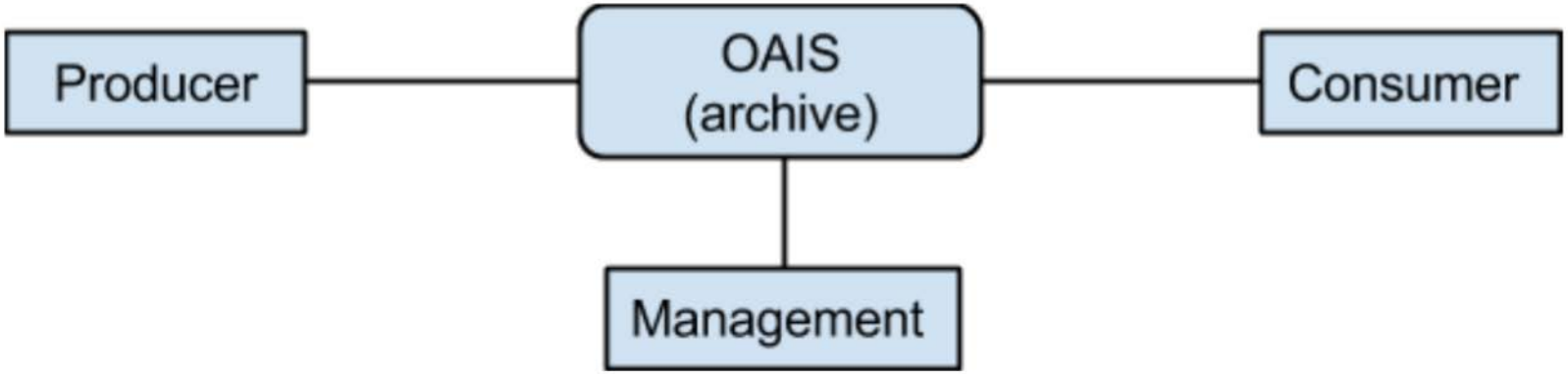
https://public.ccsds.org/Pubs/650x0m2.pdf

Draft Octobre 2020

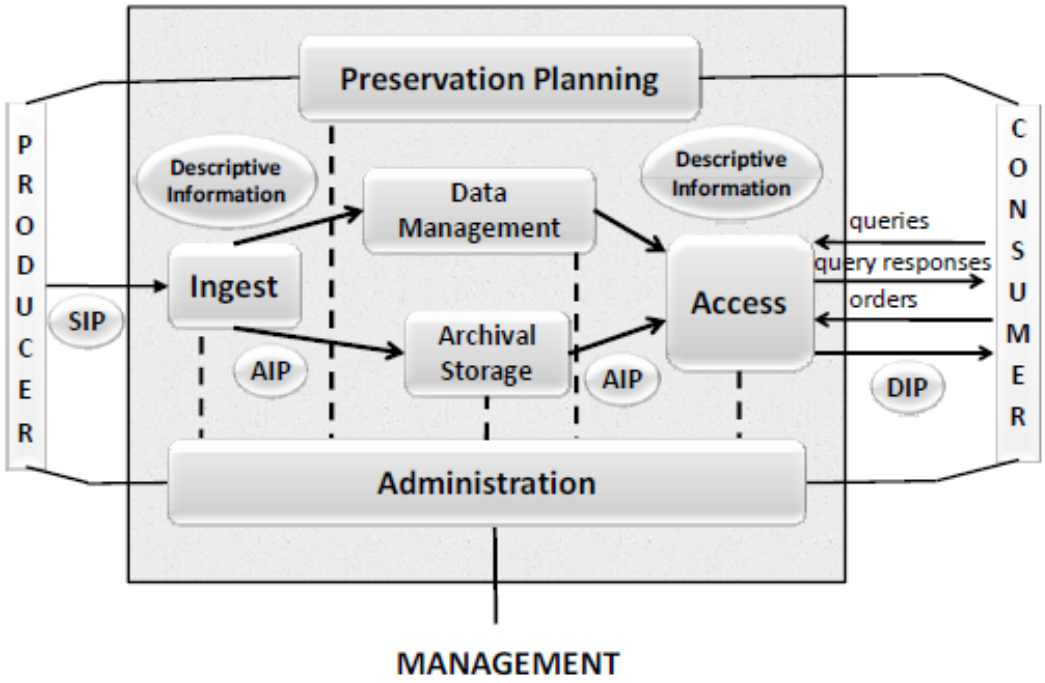
- Site en français

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

L’environnement d’une archive OAIS

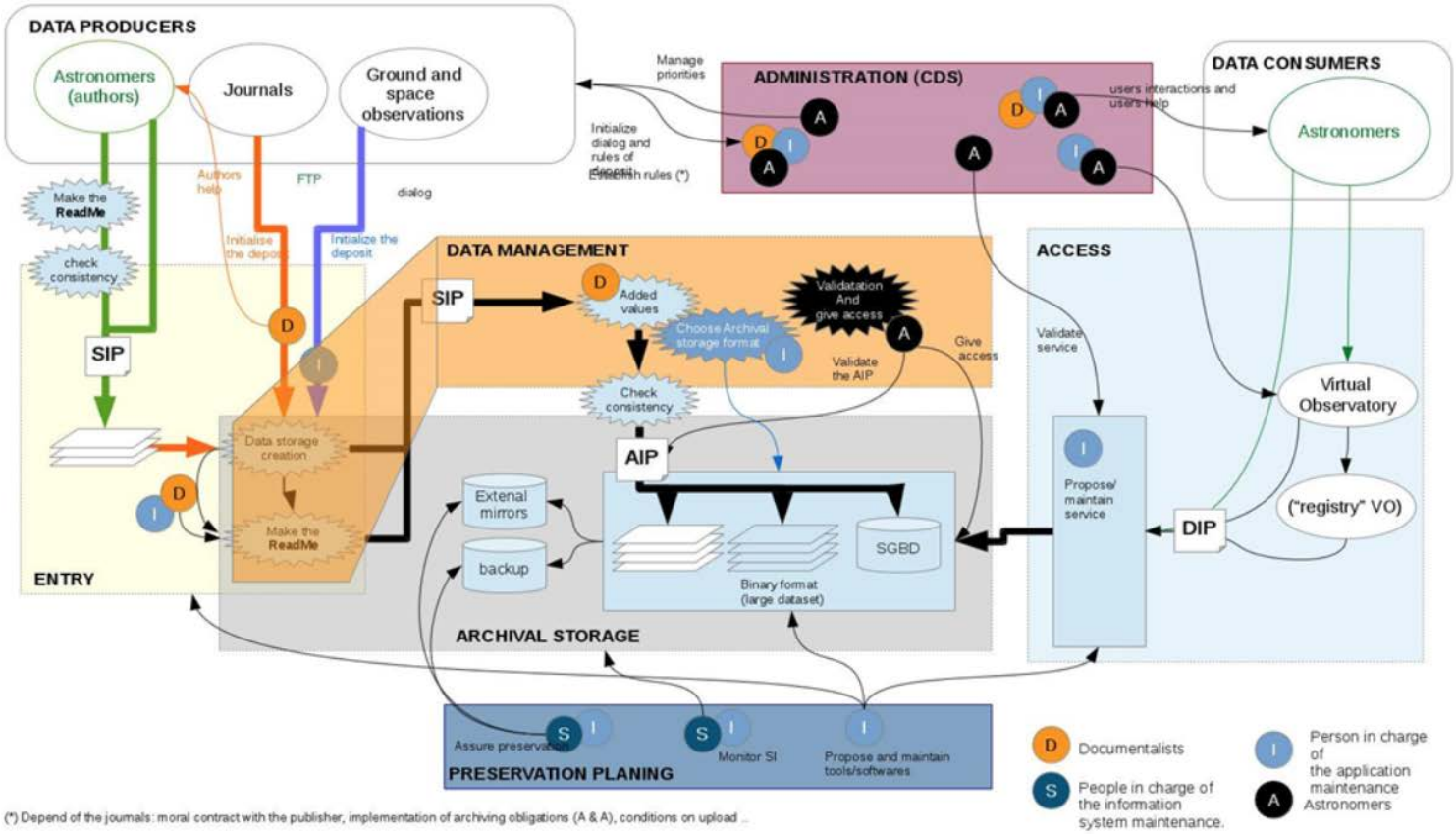


Les entités fonctionnelles de l'OAIS



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

Le pipeline de données du CDS dans le modèle OAIS



Les conséquences pour le CDS

- Description de bout en bout des process et des rôles
- Pas de modification majeure
- Des améliorations suite à l'auto-évaluation pour le DSA
  - Clarification des licences
  - Checksums des fichiers
- Le document soumis à CTS en 2018 a été accepté sans modification majeure
- Réaction très positive de nos autorités

La procédure de certification

- Soumission d'un formulaire en ligne
- Examen par deux évaluateurs
- Examen des évaluations par le Board
  - Renvoi des évaluations des évaluateurs etdes commentaires du Board aux candidats
  - Acceptation du dossier

- Les éléments du dossiers sont publics (sauf exception légitime)
- Les évaluateurs sont membres de la communauté CoreTrustSeal

Les critères du CoreTrustSeal

La certification CTS

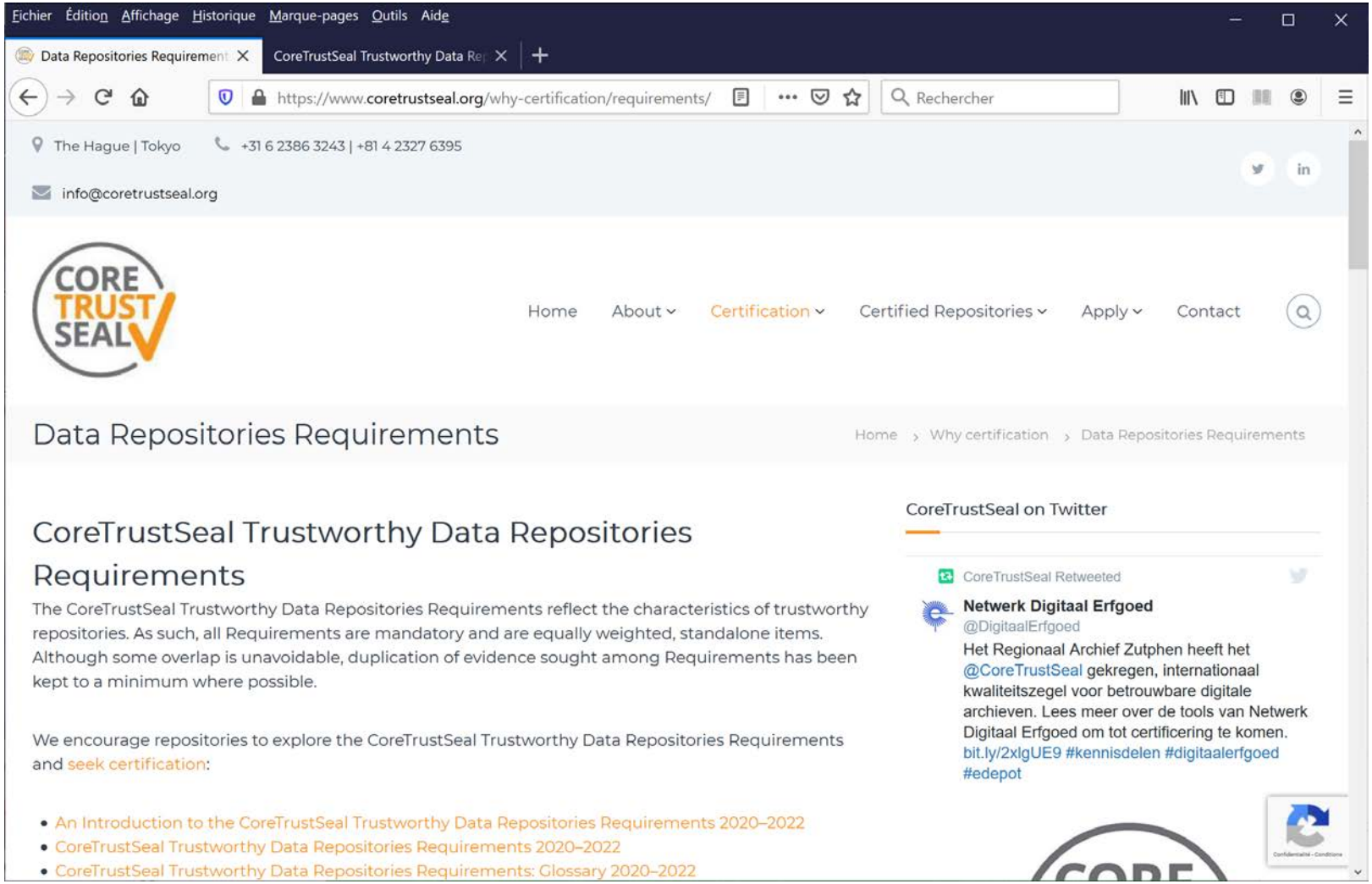
- Toute l'information est sur le site de CoreTrustSeal
  - https://www.coretrustseal.org/
- Contexte + 16 critères
  - https://www.coretrustseal.org/why-certification/requirements/
- Document pour guider les évaluateurs et les candidats
  - En cours, Extended Guidance V2.0

  https://doi.org/10.5281/zenodo.3632533
  - Traduction française par le réseau Portage

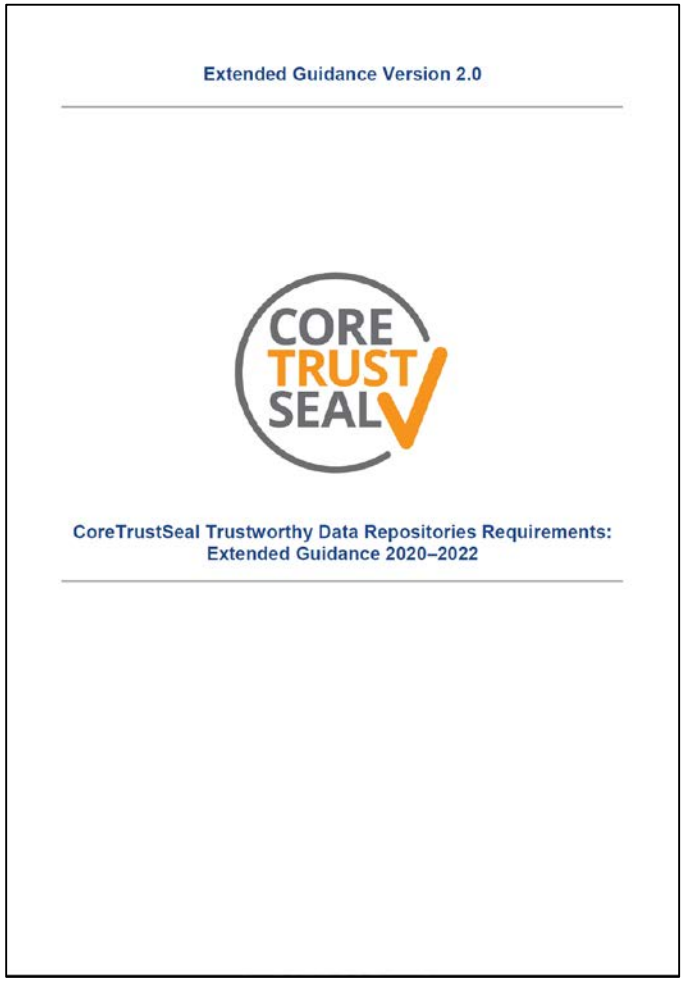
  https://doi.org/10.5281/zenodo.5046952

- « Administrative fee » 1000€

Les critères sur le site CTS



Les critères de certification CTS



- R0 - Le contexte
- 16 critères, 3 thèmes:
  - Infrastructure organisationnelle
  - Gestion des objets numériques (données et des metadonnées)
  - Technologie

Le contexte

- Type d'entrepôt
- Brève description de l'entrepôt
- Brève description de la communauté concernée
- Niveau de curation
  - Contenu en accès tel que déposé
  - Curation de base (p. ex. vérification rapide, ajout de métadonnées de base ou de documentation)
  - Curation avancée (p. ex. conversion vers de nouveaux formats, amélioration de la qualité de la documentation)
  - Curation au niveau des données
- Partenaires
- *Résumé des modifications depuis la candidature précédente (s'il y a lieu)*
- Autres informations pertinentes

Infrastructure organisationnelle

- R1 – Mission/périmètre
- R2 – Licenses
- R3 – Continuité de l’accès
- R4 – Confidentialité/éthique
- R5 – Infrastructure organisationnelle
- R6 – Conseils d’experts

Gestion des objets numériques

- R7 – Intégrité et authenticité des données
- R8 – Appréciation et sélection des données
- R9 – Procédures d'archivage documentées
- R10 – Plan de préservation
- R11 – Qualité des données
- R12 – Processus de traitement (Workflows)
- R13 – Découverte et identification des données
- R14 – Réutilisation des données

Technologie

- R15 – Infrastructure technique
- R16 – Sécurité

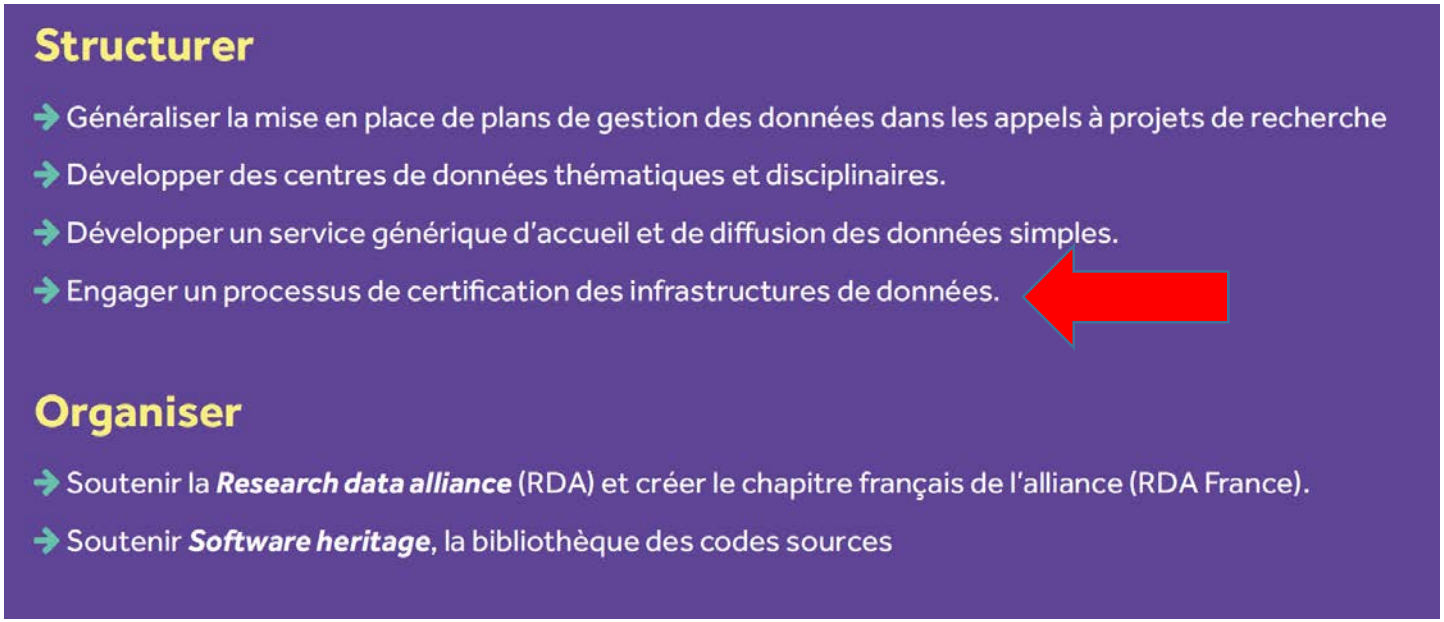
Conclusions

Pourquoi la certification?

- Quelques semaines de travail d'équipe dans le cas du CDS (tout compris)
- Evaluation – amélioration des process
  - Evaluation interne
  - Evaluation externe
- Importance croissante pour les financeurs des centres de données et des projets (DMP)
- Priorité au niveau politique en France, intérêt des organismes (CNRS, Universités)

Au niveau national: Plan national pour la Science Ouverte 2018 et 2021

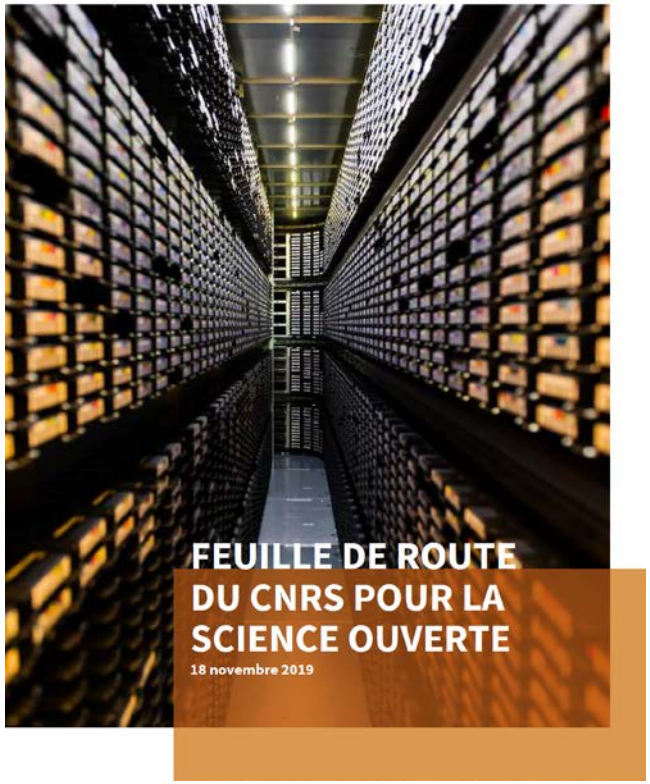
2018



2021

→ Poursuivre le processus de **certification** (*Core trust seal*) des entrepôts de données français.

Au niveau des organismes: le CNRS (1)



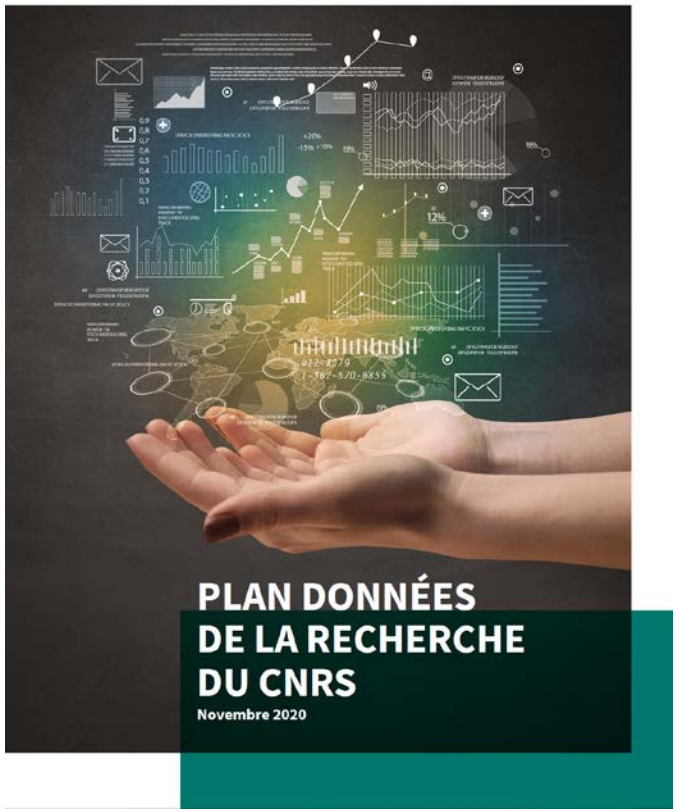
Action 3 : soutenir et accompagner les infrastructures de recherche, productrices de données, dans la définition et la mise en œuvre de politiques de données.

Le CNRS est largement engagé avec ses partenaires dans les Infrastructures de Recherche (IR) nationales et internationales, qui représentent les lieux où se créent et s'analysent les données de la recherche : instruments analytiques, infrastructures de calcul, infrastructures de données, observatoires, etc. Pour généraliser l'application des principes FAIR à toutes les disciplines, le CNRS publiera une charte des infrastructures, engageant celles-ci à respecter les pratiques FAIR et des standards de qualité, en affichant des politiques de données concertées avec les communautés scientifiques utilisatrices des infrastructures concernées. Certaines infrastructures (telles que Progedo et Humanum à l'institut des SHS (INSHS)) sont déjà bien engagées dans ce processus, d'autres sont en cours d'accompagnement telles que les IR de Chimie. Le synchrotron SOLEIL a également mis en route une politique de gestion des données. Les exemples sont multiples et devraient tendre à être généraliser. Ces développements doivent être corrélés avec les certifications (de type CoreTrustSeal) dans le cas où les infrastructures prennent elles-mêmes en charge la distribution de leurs données.

Action 4 : soutenir et accompagner des Infrastructures de données - Mettre en œuvre un service coordonné avec les instituts pour favoriser le dépôt des données pour tous les personnels des unités du CNRS

Les infrastructures de données thématiques jouent un rôle national ou international. Certaines sont inscrites sur la feuille de route nationale des infrastructures de recherche. Cela s'inscrit dans la mesure de structuration du Plan national pour la Science ouverte qui préconise de « développer des centres de données thématiques et disciplinaires ». Le CNRS continuera à soutenir ces infrastructures, et soutiendra le développement de nouveaux réservoirs et services de données thématiques. Ce soutien sera conditionné à une évaluation de leur impact, de leur adéquation aux besoins scientifiques, et de la qualité de leur gestion. Une certification CoreTrustSeal sera recherchée.

Au niveau des organismes: le CNRS (2)



p.8

- Le CNRS constituera à l'intention des chercheurs et des chercheuses un annuaire des entrepôts et des services de données existants, avec en particulier l'objectif d'aller vers la certification des entrepôts et services de données.

  Un entrepôt doit avoir un rôle de curation et de préservation des données, et les principes FAIR sont un objectif dans le contexte Science Ouverte. La certification de base *CoreTrustSeal* explicite les critères pour un entrepôt « de confiance », ce qui permet de travailler à améliorer les pratiques en se basant sur les critères, sans nécessairement aller jusqu'à soumettre un dossier de certification.

p. 11

Certification des dispositifs de prise en charge des données de la recherche (notamment le *CoreTrustSeal*[5]). La certification des entrepôts et services de données, citée comme un objectif dans le Plan National pour la Science Ouverte, permet d'assurer qu'un centre de données est « de confiance », en examinant la manière dont il met en œuvre l'ensemble de la chaîne liée aux données, de leur ingestion à leur dissémination et à leur préservation. Elle peut aussi s'entendre dans le cadre de réseaux de centres de données, par exemple ceux des Pôles de données thématiques de l'IR Data Terra[6], ou ceux de l'infrastructure européenne CLARIN[7]. Le CNRS pourra s'appuyer sur les activités de soutien à la certification mises en place par le Nœud National RDA France[8].

L’impact de la RDA

- Fusion des deux cadres de certification ‘de base’
- Clarification du paysage pour les centres de données et les agences de financement
- Deux cadres complémentaires au départ: le résultat est meilleur que chacun des originaux!
- Nombreux nouveaux candidats à la certification

Le rôle de RDA France

- La certification est une priorité
- Ateliers, présentations à la demande, etc.
- Liste de diffusion

  https://listes.services.cnrs.fr/wws/subscribe/rda-france-certification
- Groupe de Travail commun avec le collège Données du Comité pour la Science Ouverte depuis juillet 2021
  - Prise en charge des frais administratifs de certification
  - En préparation
    - Ateliers avancés
    - Outil de visualisation des dossiers de certification

RDA Global

Email - enquiries@rd-alliance.org
Web - www.rd-alliance.org
Twitter - @resdatall
LinkedIn - www.linkedin.com/in/ResearchDataAlliance
Slideshare - http://www.slideshare.net/ResearchDataAlliance

RDA FRANCE

https://rd-alliance.org/groups/rda-france

Email - contact-rdafrance@services.cnrs.fr