La certification des entrepôts de données

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

Aude Chambodut, EOST, CNRS INSU, Board CTS

Olivier Rouchon, CNRS, FAIR-IMPACT, Board CTS

Gilles Ohanessian, GT Certification

CC BY SA

La certification des entrepôts de données

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

RDA FRANCE

La certification, pourquoi?

Confiance (Trustworthiness)

- La confiance est au cœur de la Science Ouverte
- Les entrepôts sont une source majeure de données
- Il leur faut avoir la confiance
  - de leurs utilisateurs
    - Les personnes qui produisent et déposent les données
    - Les utilisateurs des données
  - de leurs autorités et de leurs financeurs

CC BY SA

Qu'est-ce qu'un entrepôt de confiance?

- Mission de fournir un accès fiable, long-terme à des ressources numériques contrôlées pour sa communauté cible, maintenant et dans le futur
- Supervision, planification et maintenance constantes
- Compréhension des menaces et risques dans les sytèmes
- **Cycle d'audit et/ou de certification régulier**

Credits: Ingrid Dillo & Hervé L'Hours

Qu'est-ce qu'un entrepôt de confiance?

- Mission de fournir un accès fiable, long-terme à des ressources numériques contrôlées pour sa communauté cible, maintenant et dans le futur
- Supervision, planification et maintenance constantes
- Compréhension des menaces et risques dans les sytèmes
- **Cycle d'audit et/ou de certification régulier** ?

Credits: Ingrid Dillo & Hervé L'Hours

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
  - 2 000 000+ requêtes/jour sur les services

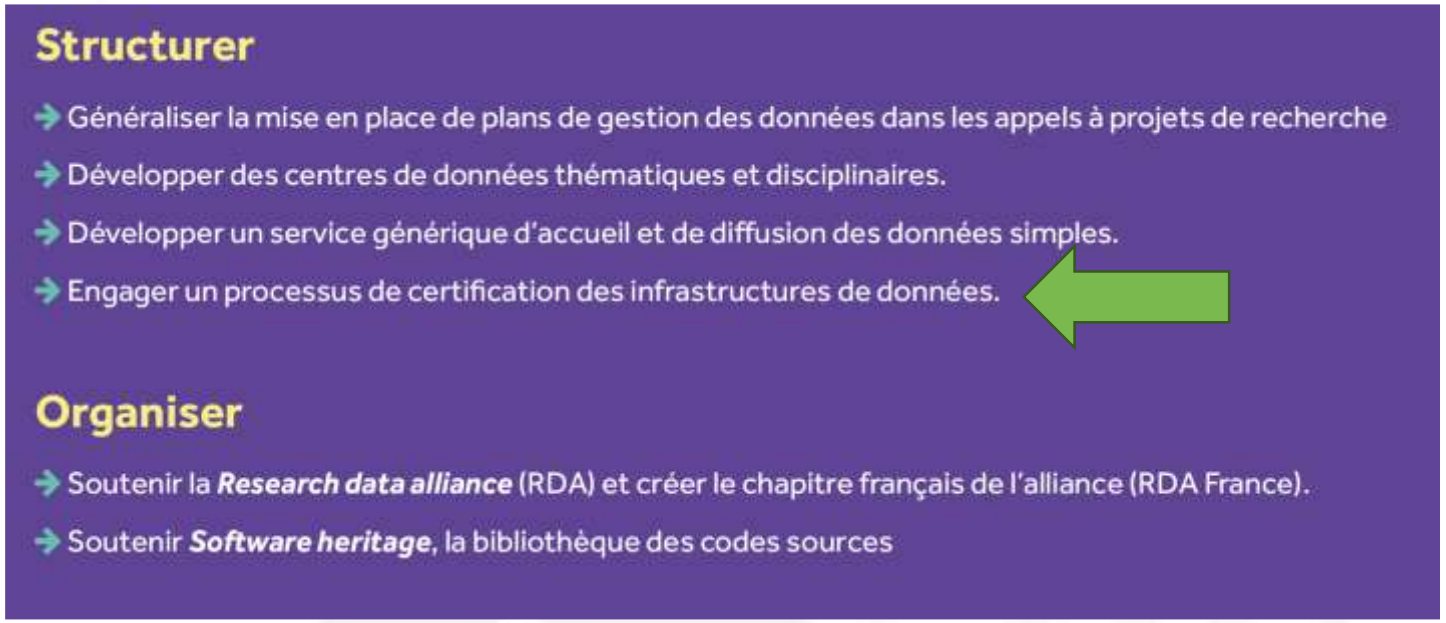
Pourquoi une certification formelle?

- 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 Plans de Gestion des Données (PGD)

La stratégie Nationale

Plan National pour la Science Ouverte

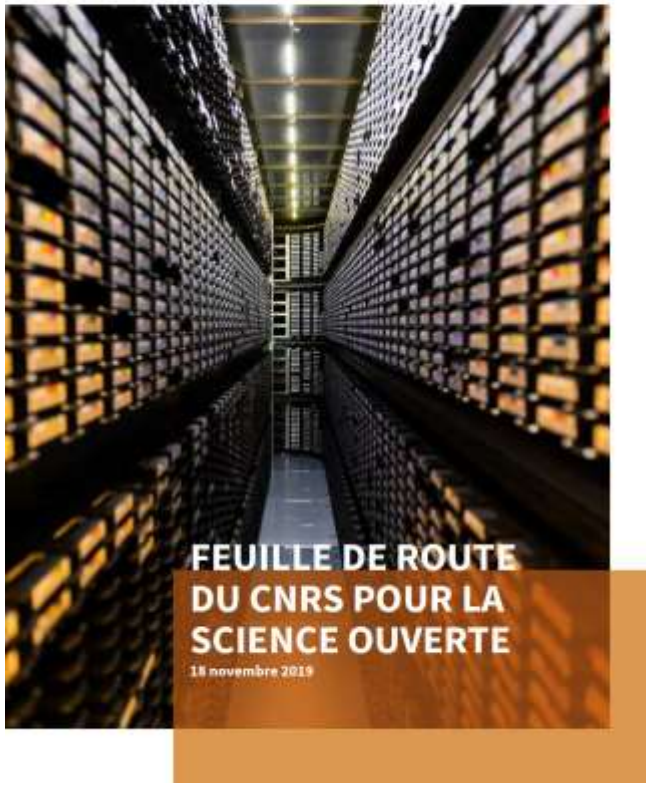
2018



2021

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

La stratégie au CNRS



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.

La stratégie au CNRS

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].

Entrepôts français actuellement certifiés CTS

- ESRF, Grenoble
- RESIF, Grenoble
- ORTOLANG, Nancy
- IDOC/IDOC-DATA, Orsay
- CDSP, Paris
- IPSL, Paris
- IFREMER SISMER, Plouzané
- CDS, Strasbourg

RDA FRANCE

Les cadres de certification ‘de base’

Un peu d’histoire…

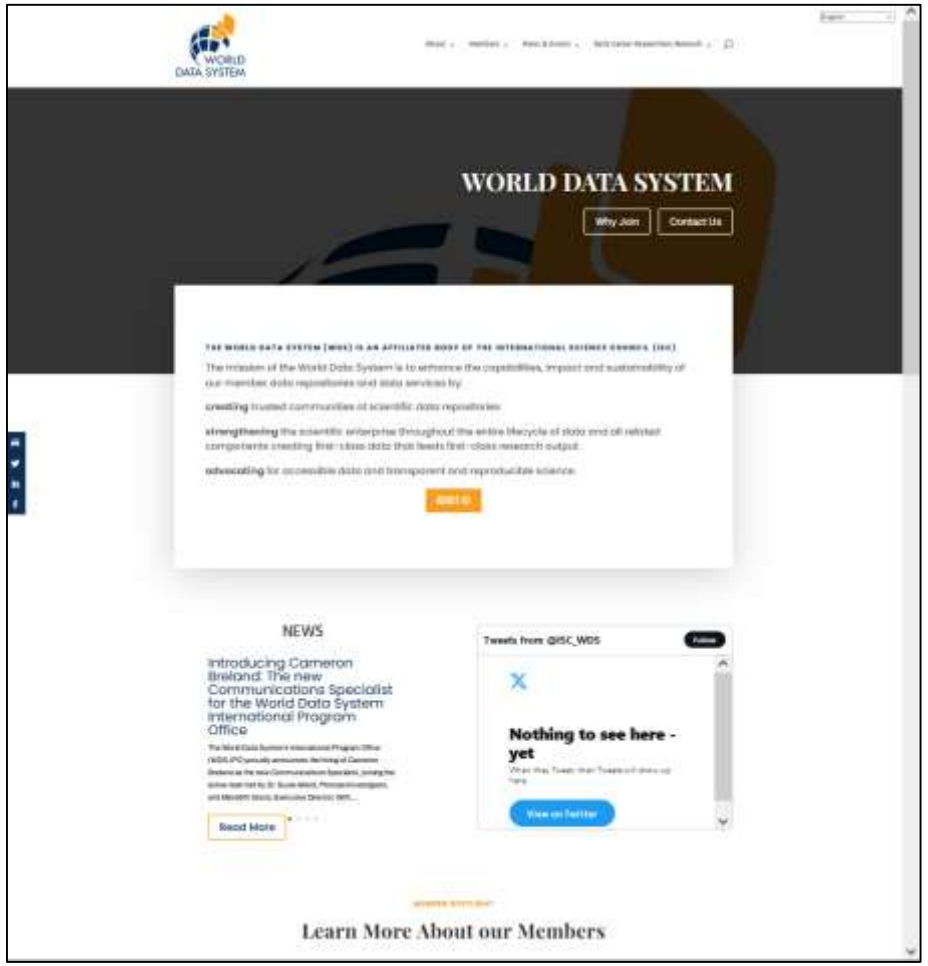
Le paysage de la certification en 2015

- 4 certifications standards disponibles

DIN 31644

ISO 16363

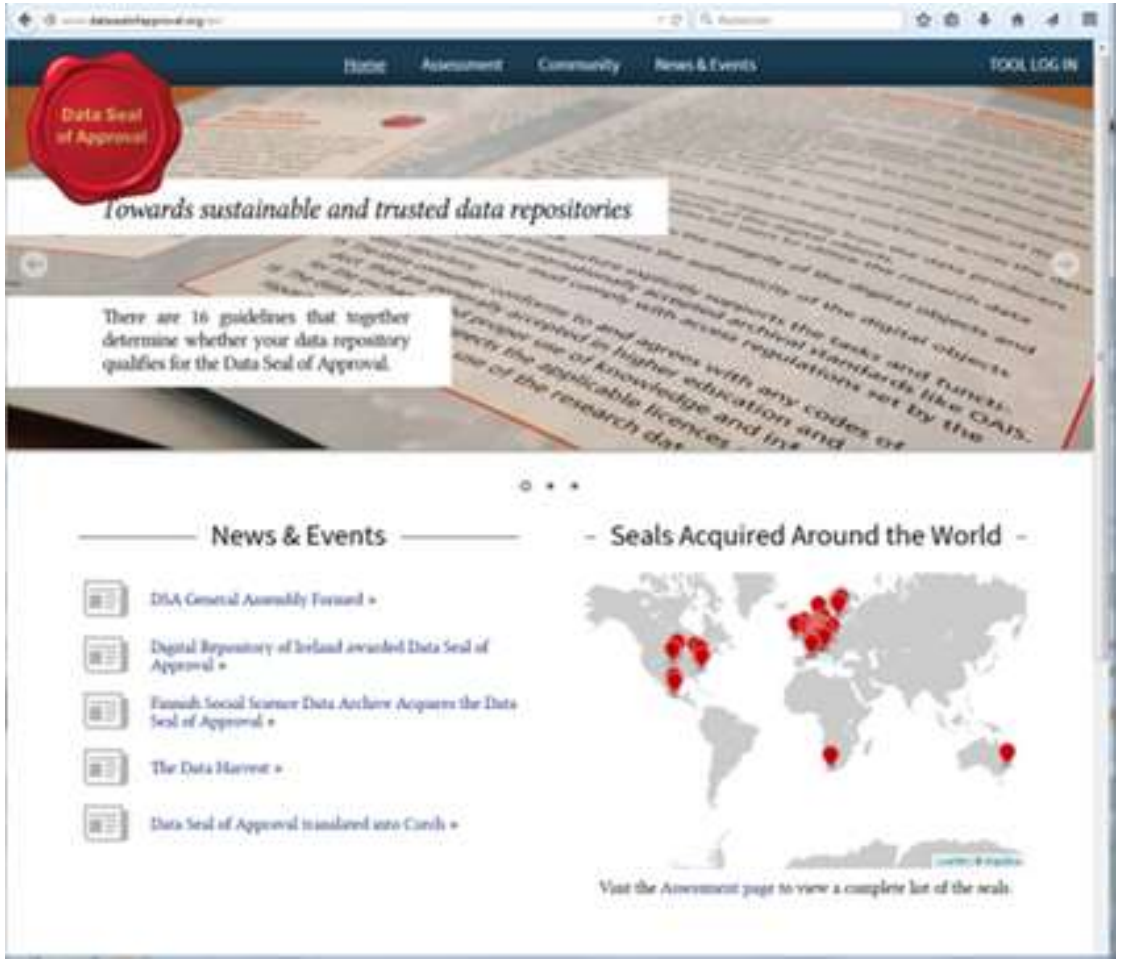
Le World Data System (WDS)



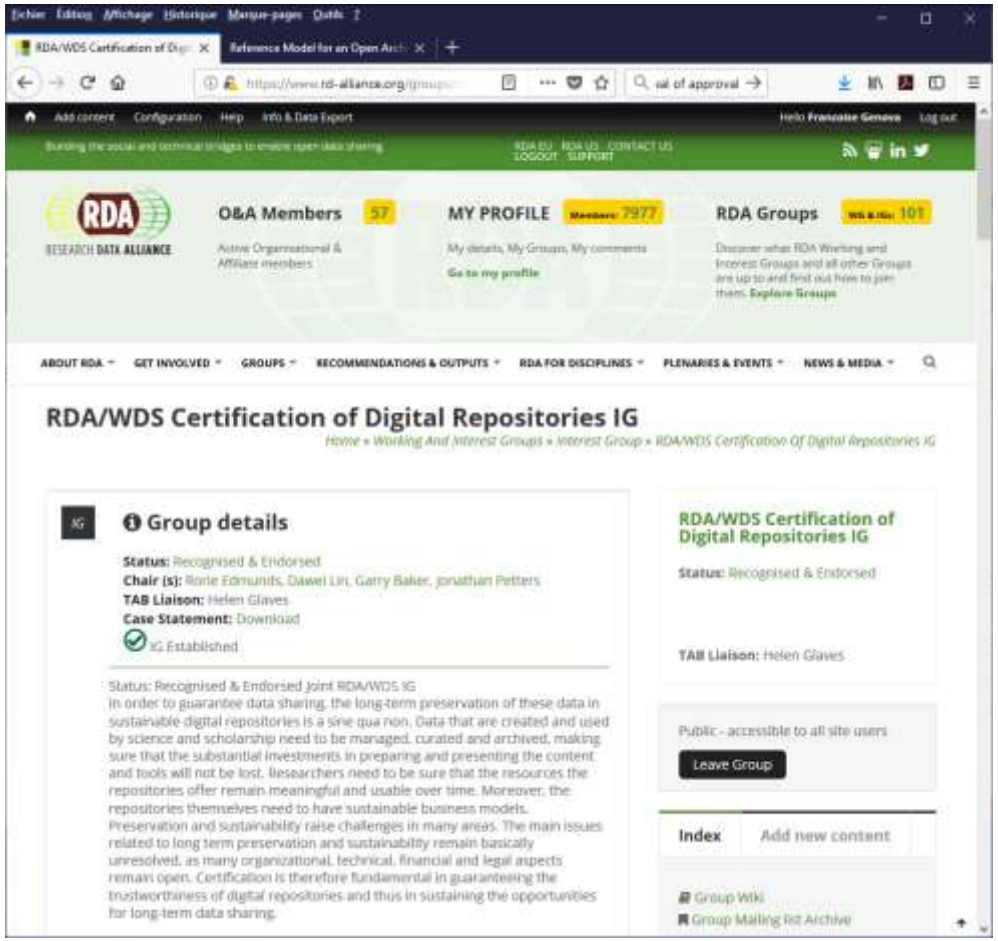
- Créé en 2008 par l'ICSU (=ISC)
- Essentiellement au départ données sur la planète (et astronomie) mais ouvert à tous
- *Promoting universal and equitable access to, and long-term stewardship of, quality-assured scientific data and data services, products, and information covering a broad range of disciplines 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 (DSA)

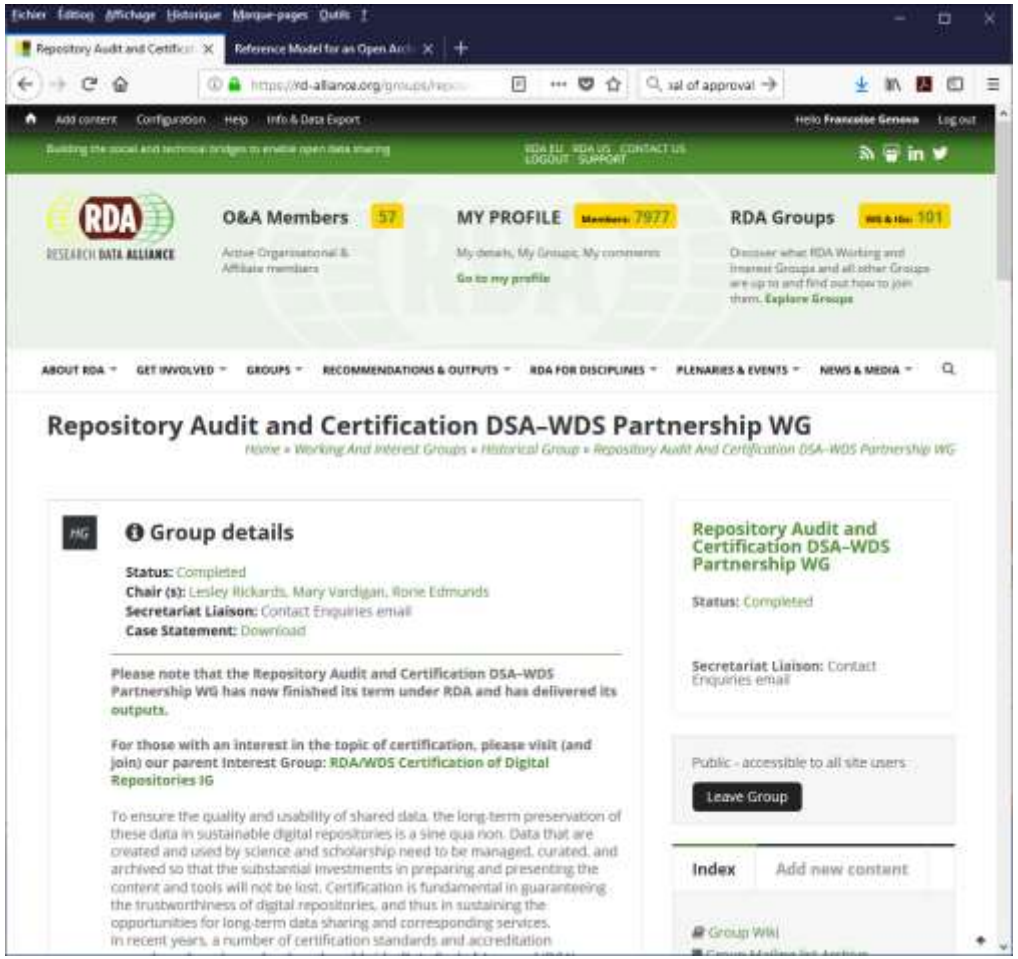
- Plutôt Humanités
- Plus dépôts de données que services
- Le CINES a été le premier centre français certifié DSA, le CDS a été le second en France et le premier centre certifié du domaine des sciences physiques (en 2014)

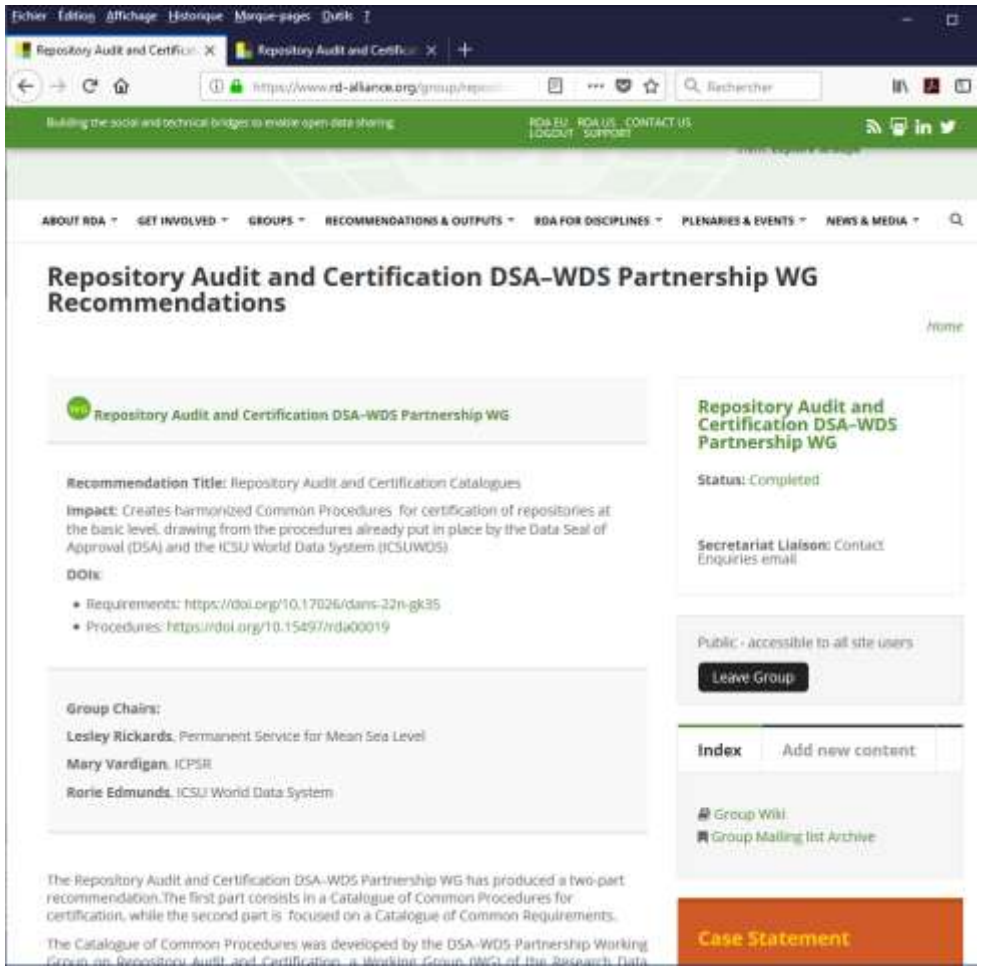


Dans la RDA, dès 2013



RDA: DSA + WDS (2016)





DSA + WDS = CoreTrustSeal (CTS), 2017

RDA FRANCE

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

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

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

Centres de données de confiance - Trustworthy Data Repositories

Credits: Mustapha Mokrane

CC BY SA

Exemple d’auto-évaluation

Le Centre de Données astronomiques de Strasbourg (CDS)

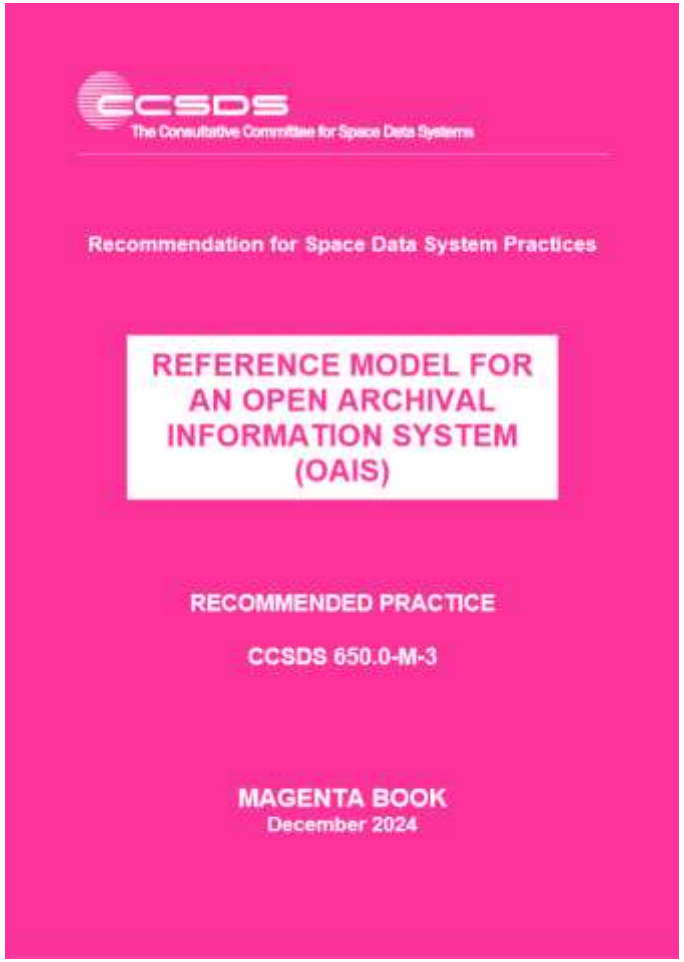
L'auto-évaluation

- 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

Les conséquences pour le CDS

- Description de bout en bout des process et des rôles
- Pas de modification majeure
- Des améliorations suite aux auto-évaluations pour DSA en 2014
  - Clarification des licences
  - Checksums des fichiers
- Le document soumis à CTS en 2018 a été accepté sans modification majeure
- Soumis en 2022 pour renouvellement en ajoutant l'entrepôt d'un second service
- Réaction très positive de nos autorités

Description des process du CDS



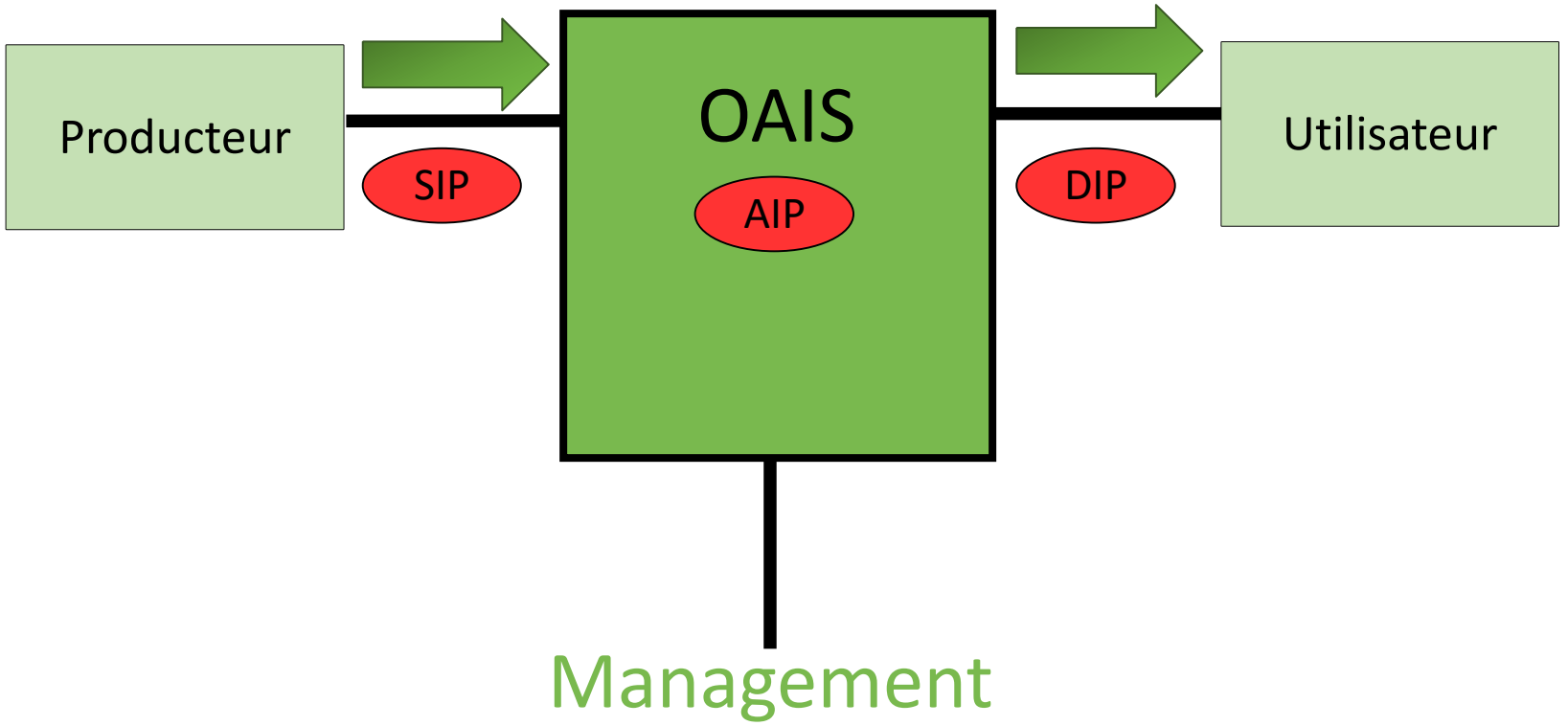
- Basé sur le modèle OAIS – Open Archive Information System Version 2
- La version actuelle: Version 3, décembre 2024

https://public.ccsds.org/Pubs/650x0m3.pdf

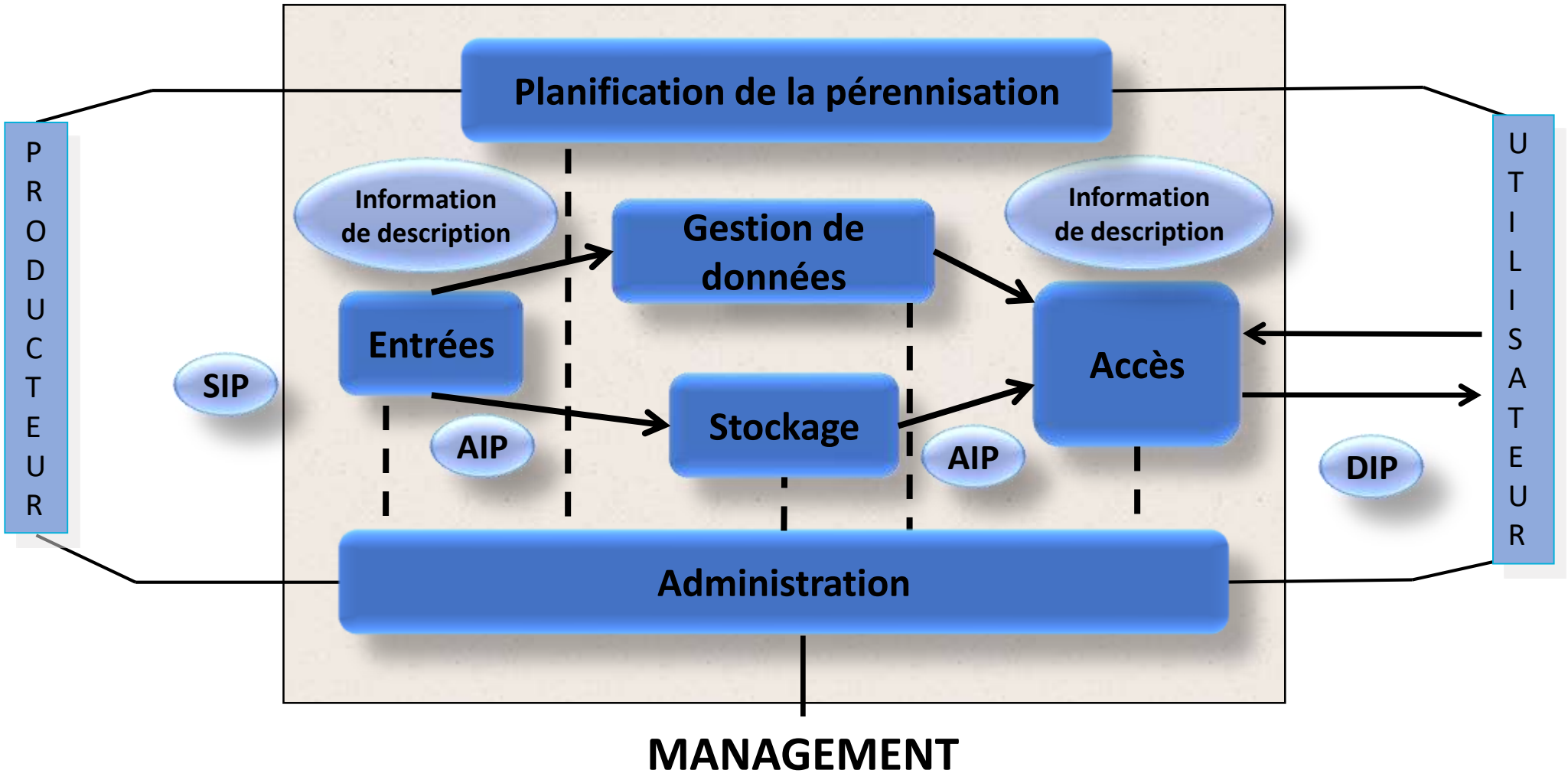
- Site en français

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

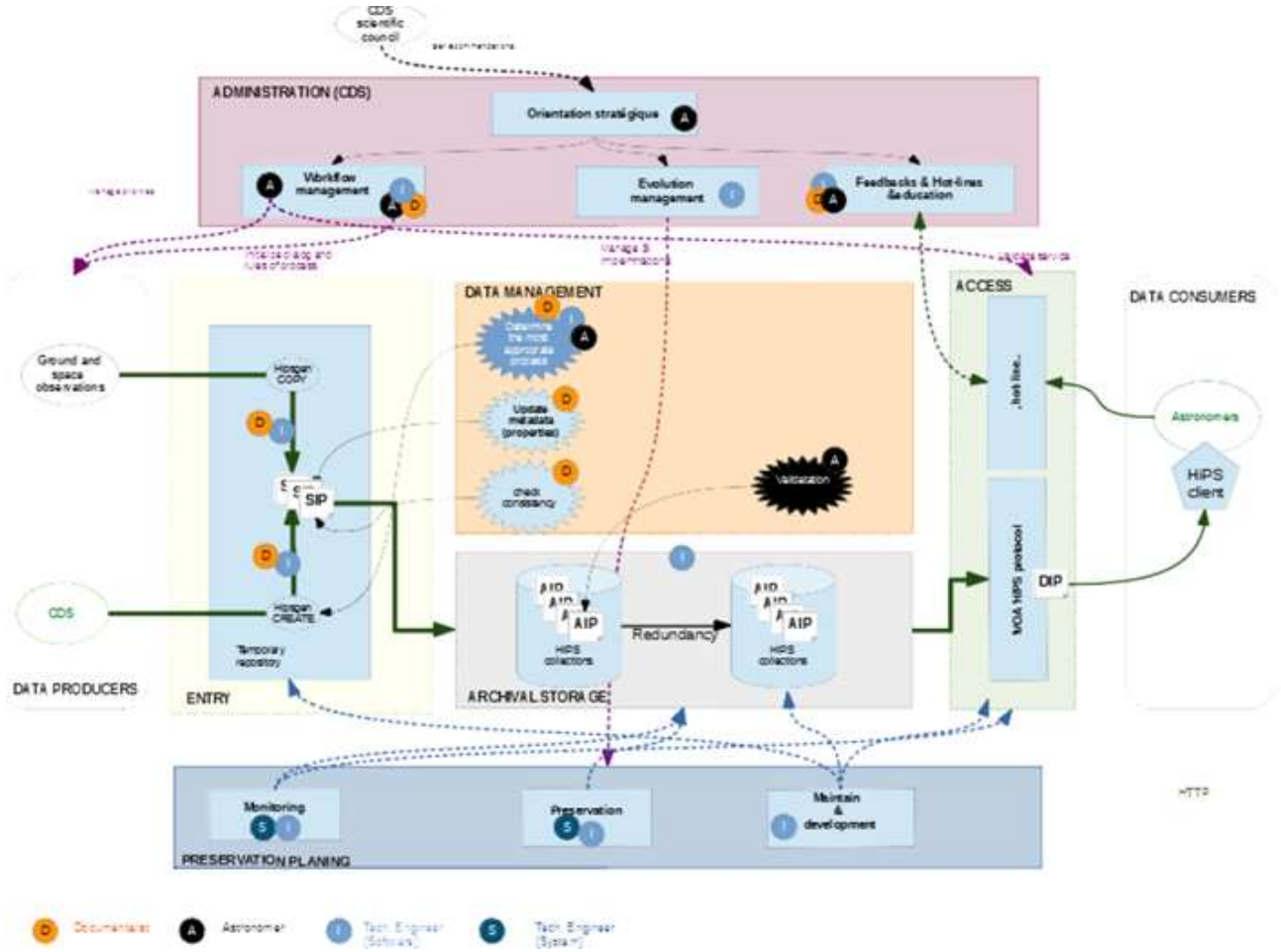
L'environnement d'une archive OAIS



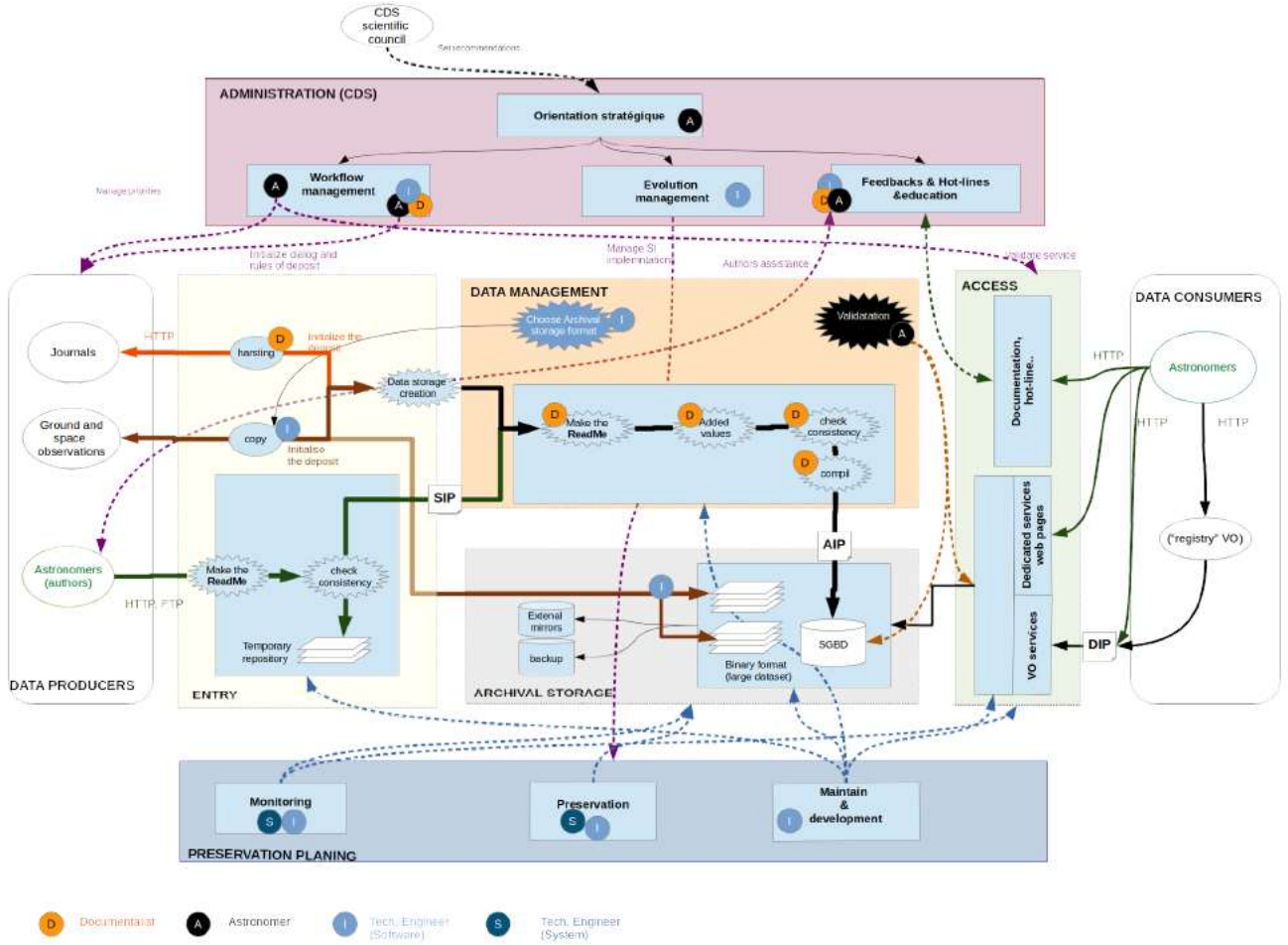
Les entités fonctionnelles de l'OAIS



Le pipeline de données pour Aladin/CDS



Le pipeline de données pour VizieR/CDS



La nouvelle version de l'OAIS

- Version 3 – publiée en Décembre 2024
  - Version 1 Janvier 2002, version 2 Juin 2012
- Principaux changements
  - Modèle conceptuel
    - Information de représentation,
    - Information de pérennisation,
    - Objectifs de préservation
  - Modèle fonctionnel
    - Planification de la préservation: veille
  - Perspectives de préservation
  - Interopérabilité de l'archive

Le cas de Dataverse

- Guide Dataverse pour la certification CTS
  - https://dataverse.org/cts-guide
  - Publié en Mars 2021
  - Basé sur la version des critères 2020-2022
  - Combine
    - Eléments de réponses basé sur la technologie et le processus de l'outil
    - Exemples de réponses d'entrepôts Dataverse certifiés (Tilburg, QDR, dataverseNO, etc.)
  - Certains points organisationnels ne sont pas détaillés
    - Mission
    - Continuité de service
  - Pertinent pour la partie technique

La procédure de certification

La procédure d'évaluation (1)

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

La procédure d'évaluation (2)

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

- Pour devenir membre du WDS
  - Il faut obtenir la certification CTS…
  - et remplir quelques conditions supplémentaires
- Renouvellement de la certification tous les 3 ans

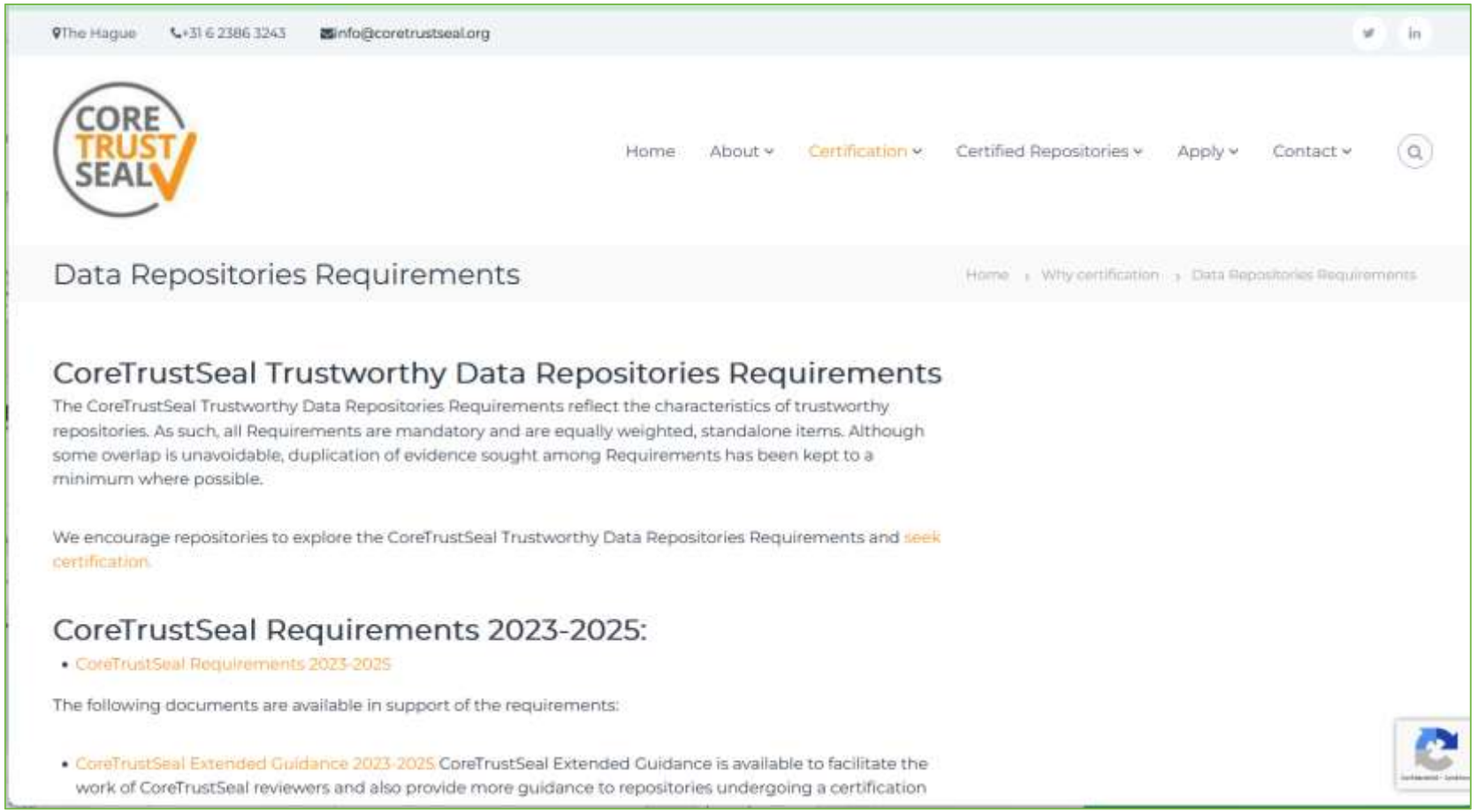
CC BY SA

Les critères du CoreTrustSeal

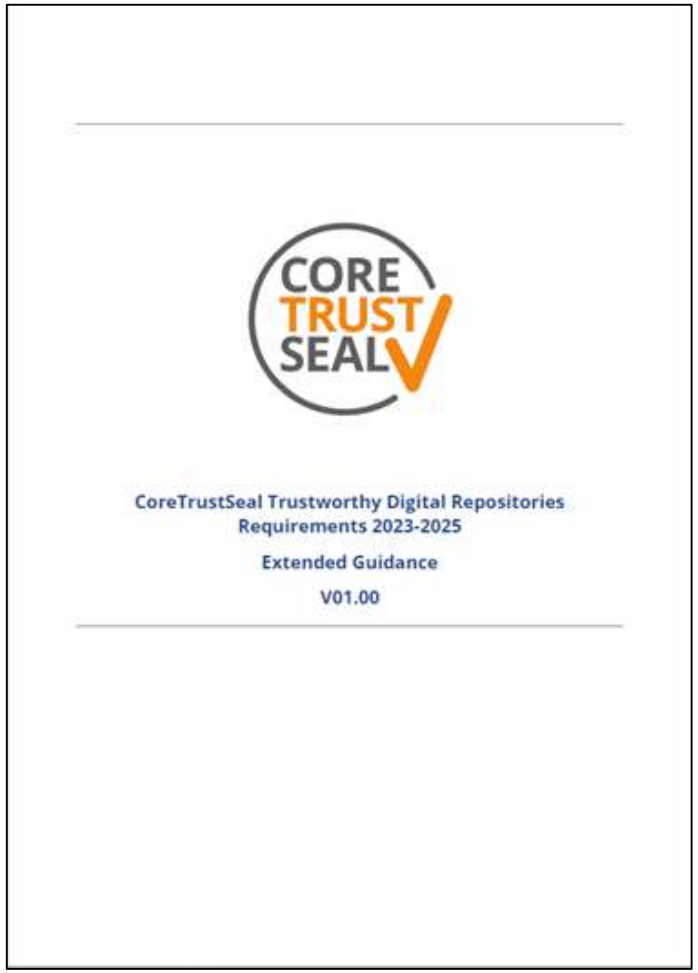
La certification CTS

- 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
  - Extended Guidance 2020-2022 > 2023-2025
    https://doi.org/10.5281/zenodo.7051096
  - Glossaire
    https://doi.org/10.5281/zenodo.7051125
  - Changements V2.0 > V3.0
    https://doi.org/10.5281/zenodo.7051237
- Les entrepôts qui ont soumis un dossier selon les anciens critères restent dans ce cadre
- « Administrative fee » 3000€

Les critères sur le site CTS

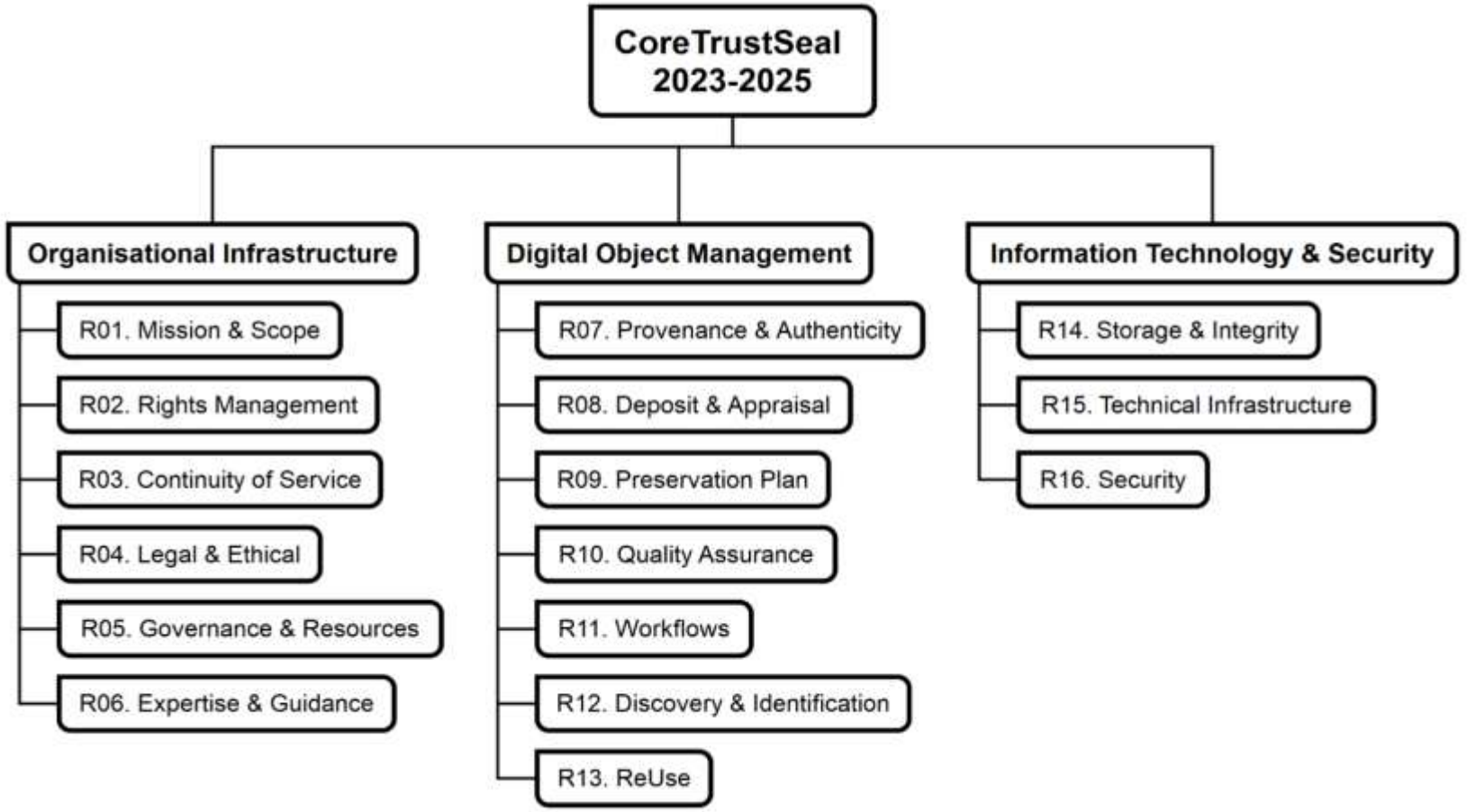


Les critères de certification CTS



- R0 - Le contexte
- 16 critères, 3 thèmes:
  - Infrastructure organisationnelle
  - Gestion des objets numériques (données et métadonnées)
  - Technologie de l'information et sécurité

Les critères 2023-2025

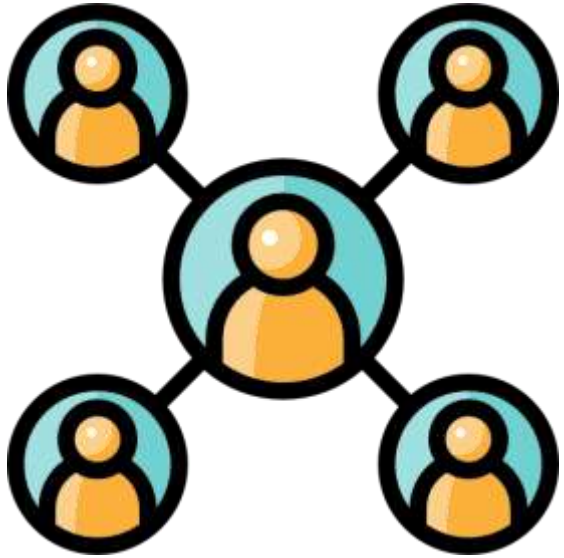


Le contexte

- Identifiant Re3data
- 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 – Gestion des droits
- R3 – Continuité de service
- R4 – Légalité/éthique
- R5 – Gouvernance et ressources
- R6 – Conseils d'experts

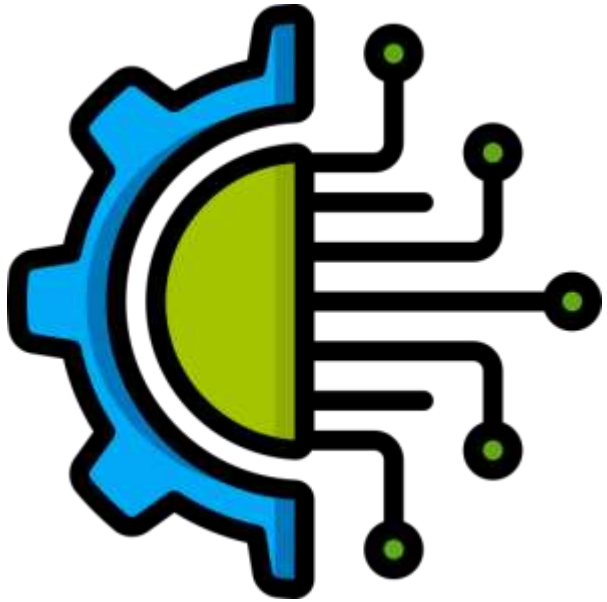


Gestion des objets numériques

- R7 – Provenance et authenticité des données
- R8 – Appréciation et sélection des données
- R9 – Assurance qualité
- R10 – Plan de préservation
- R11 – Processus de traitement (Workflows)
- R12 – Découverte et identification des données
- R13 – Réutilisation des données

Technologie

- R14 – Stockage et intégrité
- R15 – Infrastructure technique
- R16 – Sécurité



Révision des critères 2026-2028

Revue

- Revue des dossiers de candidatures des trois dernières années
- Proposition de changements et additions

Consultation

- Sondage de la communauté
- Revue des retours
- Incorporation des retours au projet de révision
- Revue par le Board et approbation de la nouvelle version

Publication

- Publication de la version finale
- Correction des dernières erreurs par la communauté et le Board
- Publication des critères 2026-2028

RDA FRANCE

Conclusions

CC BY SA

Pourquoi la certification?

- Quelques semaines de travail d'équipe dans la plupart des cas (tout compris)
  - Beaucoup plus pour d'autres mais ça continue à valoir le coup!
- Evaluation – amélioration des process
  - Auto-évaluation
  - Evaluation externe
- Importance croissante pour les financeurs des centres de données et des projets (PGD)
- Priorité au niveau politique en France, intérêt des organismes (CNRS, Universités)

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!
- Création de CoreTrustSeal
- Nombreux nouveaux candidats à la certification

CC BY SA

Le rôle de RDA France

- La certification est une priorité – aussi parce que la réflexion sur les critères permet d'améliorer la manière dont on gère ses données, même si on ne va pas jusqu'à candidater à CoreTrustSeal
- 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
  - Outil de visualisation et de partage des dossiers de certification
    https://crusoe.ouvrirlascience.fr/
- Ateliers, présentations à la demande, etc.
  - Plus de 300 personnes ont suivi les ateliers
- Démarrage de la mise en réseau des entrepôts certifiés et « sur le chemin de la certification »
  - Ateleirs pendant les Journées RDA France 2023 et 2024
- Liste de diffusion
  https://listes.services.cnrs.fr/wws/subscribe/rda-france-certification

RDA FRANCE

RDA Global

Contact - https://www.rd-alliance.org/contact/

Web - www.rd-alliance.org

Twitter - @resdatall

LinkedIn - www.linkedin.com/in/ResearchDataAlliance

Slideshare - http://www.slideshare.net/ResearchDataAlliance

RDA FRANCE

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

Email - contact-rdafrance@services.cnrs.fr