



HAL
open science

La certification des entrepôts de données

Françoise Genova, Aude Chambodut, Olivier Rouchon, Gilles Ohanessian

► **To cite this version:**

Françoise Genova, Aude Chambodut, Olivier Rouchon, Gilles Ohanessian. La certification des entrepôts de données. Printemps de la Donnée 2024, INRAE; Université Haute-Alsace; Université de Strasbourg; INSA; PNDB; AgroParisTech; Université de Lille; Sorbonne Université; Data Terra, Mar 2024, Strasbourg, France. hal-04680311

HAL Id: hal-04680311

<https://hal.inrae.fr/hal-04680311v1>

Submitted on 28 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License



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



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



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

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 systèmes
- **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 systèmes
- **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
 - Créé 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

Structurer

- Généraliser la mise en place de plans de gestion des données dans les appels à projets de recherche
- Développer des centres de données thématiques et disciplinaires.
- Développer un service générique d'accueil et de diffusion des données simples.
- Engager un processus de certification des infrastructures de données. ←

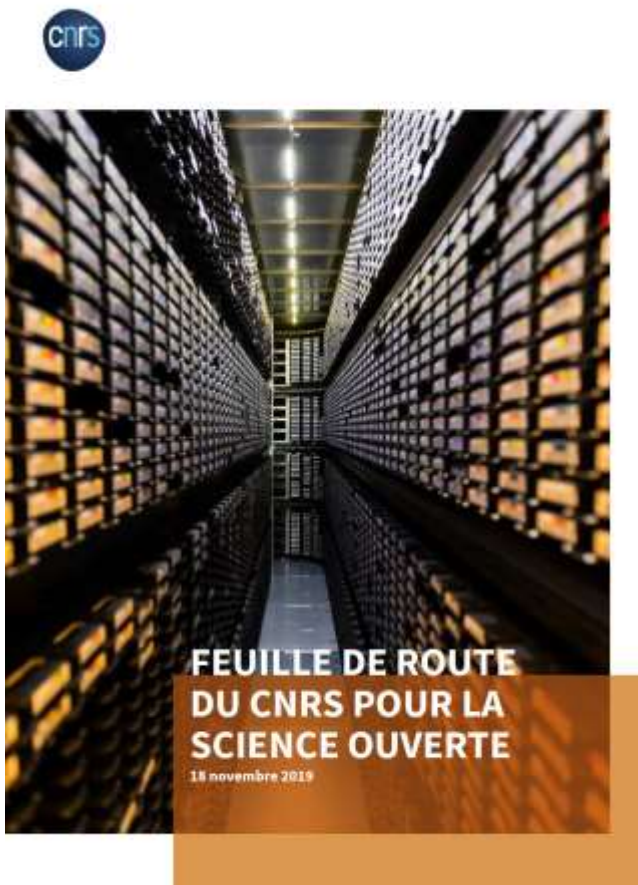
Organiser

- Soutenir la *Research data alliance* (RDA) et créer le chapitre français de l'alliance (RDA France).
- Soutenir *Software heritage*, la bibliothèque des codes sources

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. Ces exemples sont multiples et devraient tendre à être généralisés. 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 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.



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 Terra⁶, ou ceux de l'infrastructure européenne CLARIN⁷. Le CNRS pourra s'appuyer sur les activités de soutien à la certification mises en place par le Nœud National RDA France⁸.



Entrepôts français actuellement certifiés CTS

- [IDOC/IDOC-DATA](#), Orsay
- [IFREMER SISMER](#), Plouzané
- [ORTOLANG](#), Nancy
- [ESRF](#), Grenoble
- [CDSP](#), Paris
- [IPSL](#), Paris
- [CDS](#), Strasbourg





Les cadres de certification 'de base'

Un peu d'histoire...



Le paysage de la certification en 2015

4 certifications standards disponibles



DIN 31644

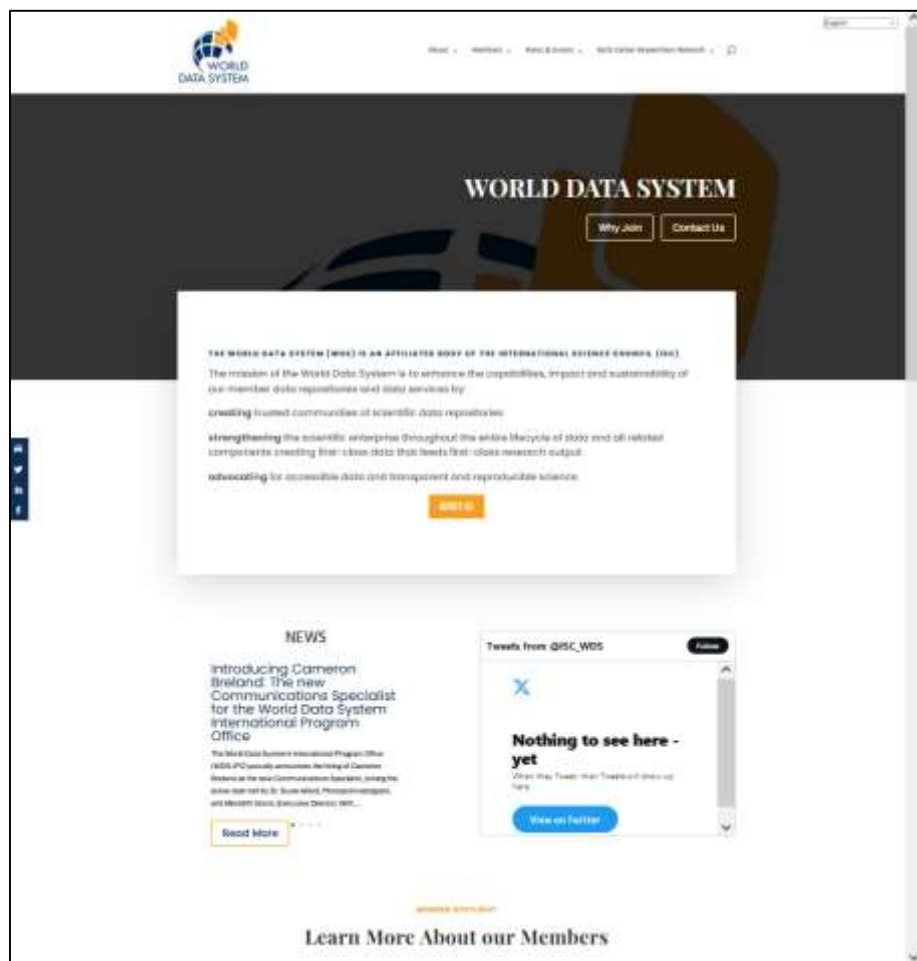


WORLD DATA SYSTEM



ISO 16363

Le World Data System (WDS)

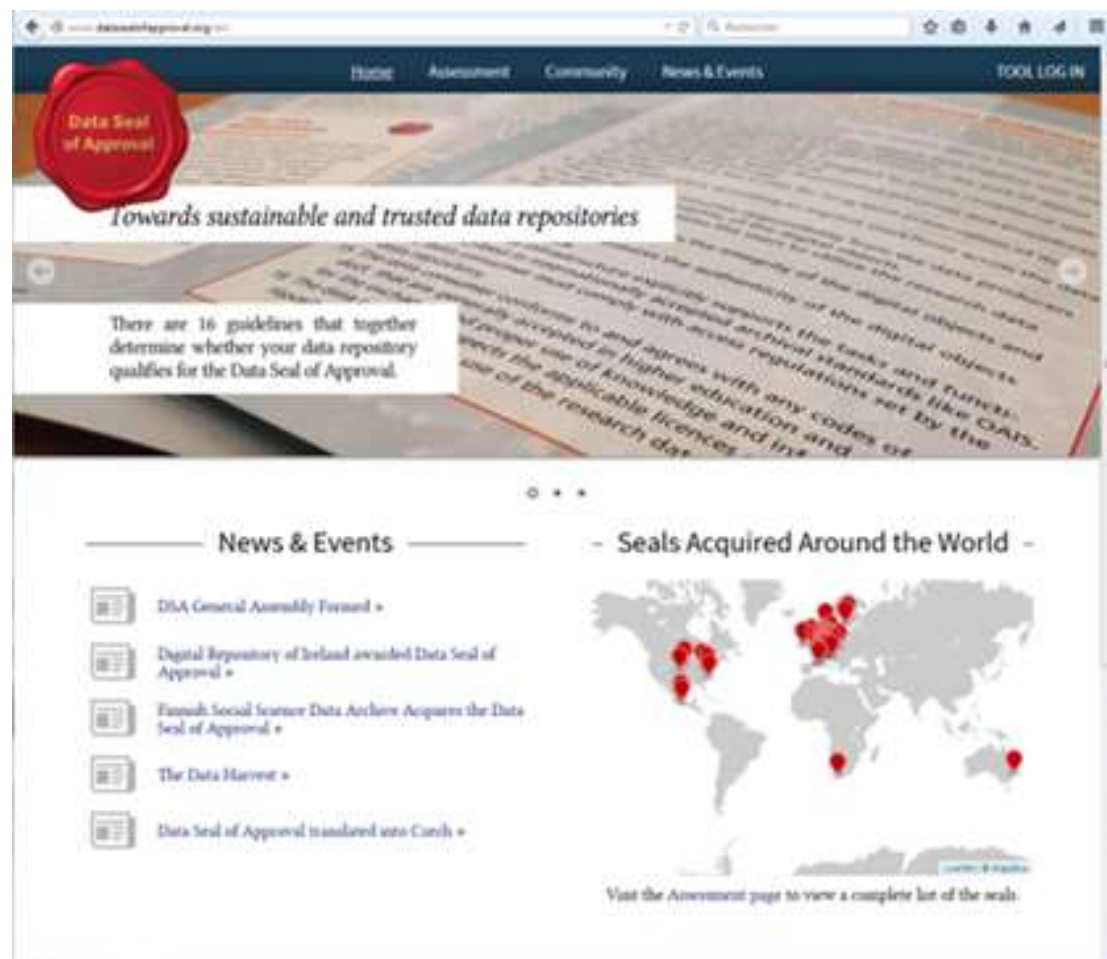


- Créé en 2008 par l'ICSU (=ISC)
- Essentiellement au départ données sur la planète (et astronomie) mais ouvert à tous
- *Promoting universal and equitable access to, and long-term stewardship of, quality-assured scientific data and data services, products, and information covering a broad range of disciplines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

A screenshot of a web browser displaying the RDA website. The page is titled "RDA/WDS Certification of Digital Repositories IG". The browser's address bar shows the URL "https://www.rd-alliance.org/groups...". The website header includes navigation links like "Add content", "Configuration", "Help", "Info & Data Export", and "Hello Françoise Genève". The main content area features a navigation menu with "ABOUT RDA", "GET INVOLVED", "GROUPS", "RECOMMENDATIONS & OUTPUTS", "RDA FOR DISCIPLINES", "PLENARIES & EVENTS", and "NEWS & MEDIA". The main heading is "RDA/WDS Certification of Digital Repositories IG" with a breadcrumb trail: "Home > Working And Interest Groups > Interest Group > RDA/WDS Certification Of Digital Repositories IG". The page is divided into two columns. The left column, titled "Group details", lists: "Status: Recognised & Endorsed", "Chair (s): Ronie Edmunds, Dawei Lin, Garry Baker, jonathan Petters", "TAB Liaison: Helen Graves", "Case Statement: Download", and "IG Established". Below this is a paragraph of text: "Status: Recognised & Endorsed Joint RDA/WDS IG. In order to guarantee data sharing, the long-term preservation of these data in sustainable digital repositories is a sine qua non. Data that are created and used by science and scholarship need to be managed, curated and archived, making sure that the substantial investments in preparing and presenting the content and tools will not be lost. Researchers need to be sure that the resources the repositories offer remain meaningful and usable over time. Moreover, the repositories themselves need to have sustainable business models. Preservation and sustainability raise challenges in many areas. The main issues related to long term preservation and sustainability remain basically unresolved, as many organizational, technical, financial and legal aspects remain open. Certification is therefore fundamental in guaranteeing the trustworthiness of digital repositories and thus in sustaining the opportunities for long-term data sharing." The right column, titled "RDA/WDS Certification of Digital Repositories IG", lists: "Status: Recognised & Endorsed", "TAB Liaison: Helen Graves", "Public - accessible to all site users", and a "Leave Group" button. At the bottom of the right column, there are links for "Index" and "Add new content", and a list of "Group Wiki" and "Group Mailing List Archive".



RDA: DSA + WDS (2016)

Repository Audit and Certification DSA-WDS Partnership WG

Home » Working And Interest Groups » Historical Group » Repository Audit And Certification DSA-WDS Partnership WG

Group details

Status: Completed
 Chair (s): Lesley Rickards, Mary Vardigan, Rorie Edmunds
 Secretariat Liaison: Contact Enquiries email
 Case Statement: Download

Please note that the Repository Audit and Certification DSA-WDS Partnership WG has now finished its term under RDA and has delivered its outputs.

For those with an interest in the topic of certification, please visit (and join) our parent Interest Group: [RDA/WDS Certification of Digital Repositories IG](#)

To ensure the quality and usability of shared data, the long-term preservation of these data in sustainable digital repositories is a sine qua non. Data that are created and used by science and scholarship need to be managed, curated, and archived so that the substantial investments in preparing and presenting the content and tools will not be lost. Certification is fundamental in guaranteeing the trustworthiness of digital repositories, and thus in sustaining the opportunities for long-term data sharing and corresponding services. In recent years, a number of certification standards and accreditation

Repository Audit and Certification DSA-WDS Partnership WG

Status: Completed

Secretariat Liaison: Contact Enquiries email

Public - accessible to all site users
 Leave Group

Index Add new content

Group Wiki

Repository Audit and Certification DSA-WDS Partnership WG Recommendations

Home

Repository Audit and Certification DSA-WDS Partnership WG

Recommendation Title: Repository Audit and Certification Catalogues

Impact: Creates harmonized Common Procedures for certification of repositories at the basic level, drawing from the procedures already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSUWDS)

DOI:

- Requirements: <https://doi.org/10.17026/dans-22n-gk35>
- Procedures: <https://rdol.org/10.15497/rd00019>

Group Chairs:

- Lesley Rickards, Permanent Service for Mean Sea Level
- Mary Vardigan, ICPSR
- Rorie Edmunds, ICSU World Data System

The Repository Audit and Certification DSA-WDS Partnership WG has produced a two-part recommendation. The first part consists in a Catalogue of Common Procedures for certification, while the second part is focused on a Catalogue of Common Requirements.

The Catalogue of Common Procedures was developed by the DSA-WDS Partnership Working Group on Repository Audit and Certification, a Working Group (WG) of the Research Data

Status: Completed

Secretariat Liaison: Contact Enquiries email

Public - accessible to all site users
 Leave Group

Index Add new content

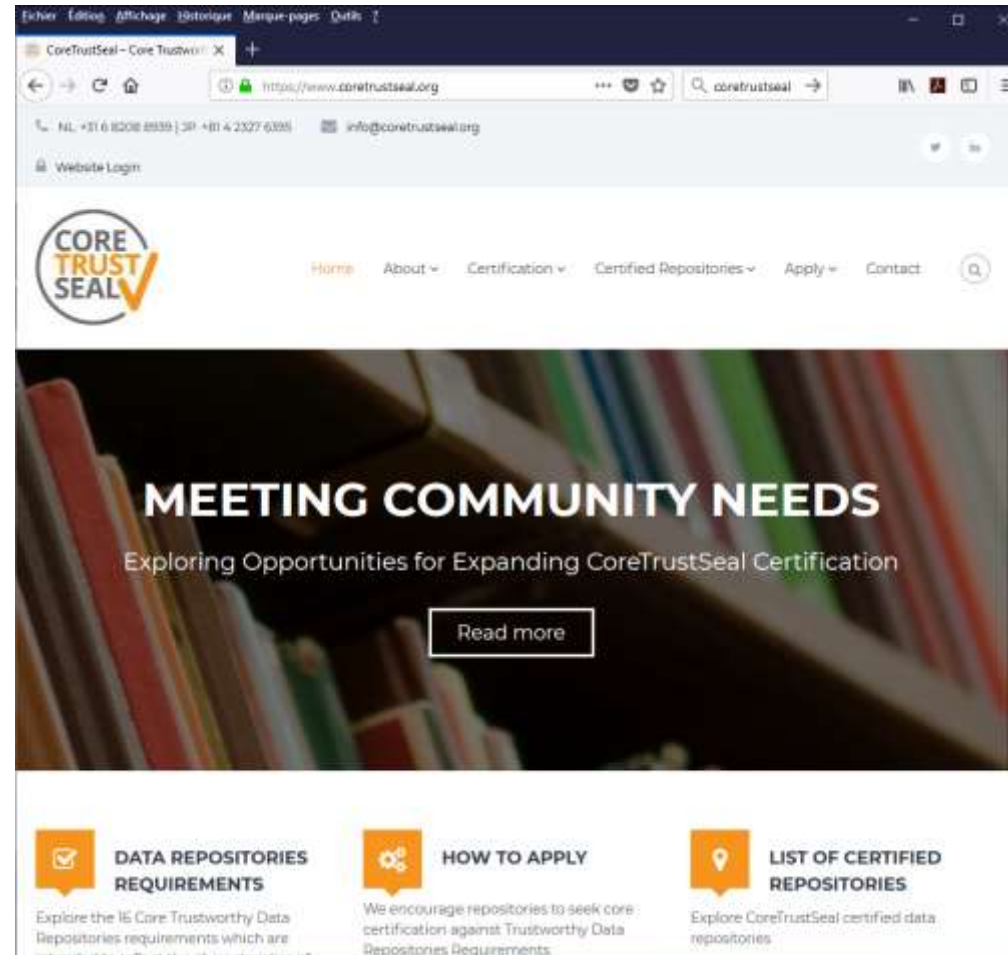
Group Wiki
 Group Mailing list Archive

Case Statement





DSA + WDS = CoreTrustSeal (CTS), 2017





Centres de données de confiance - Trustworthy Data Repositories

Credits: Mustapha Mokrane



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

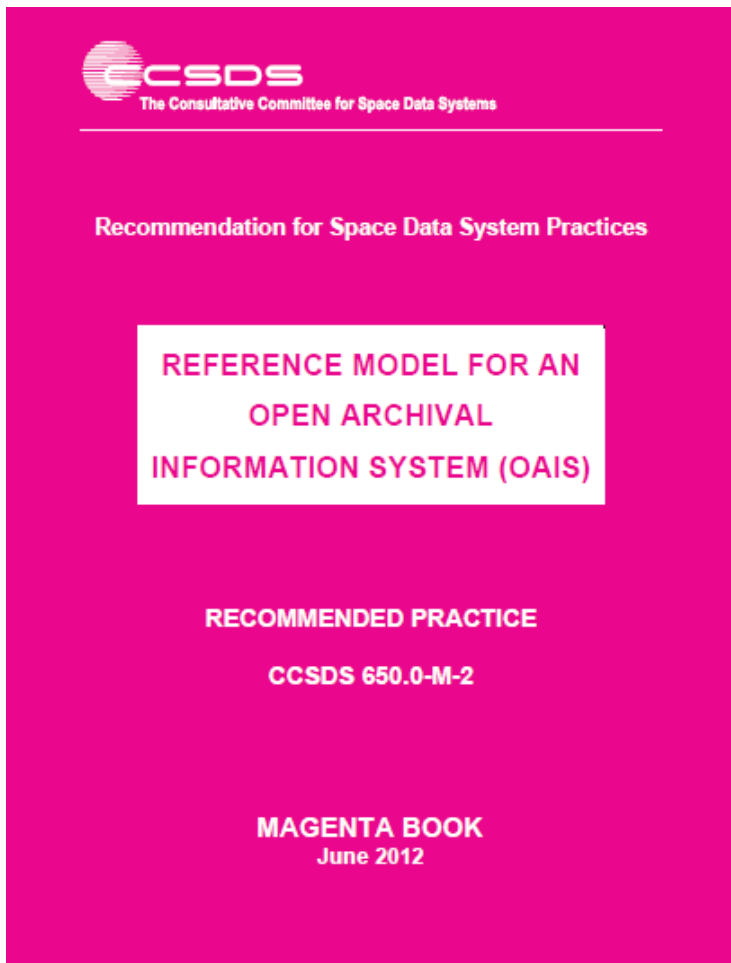




Les conséquences pour le CDS

- 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

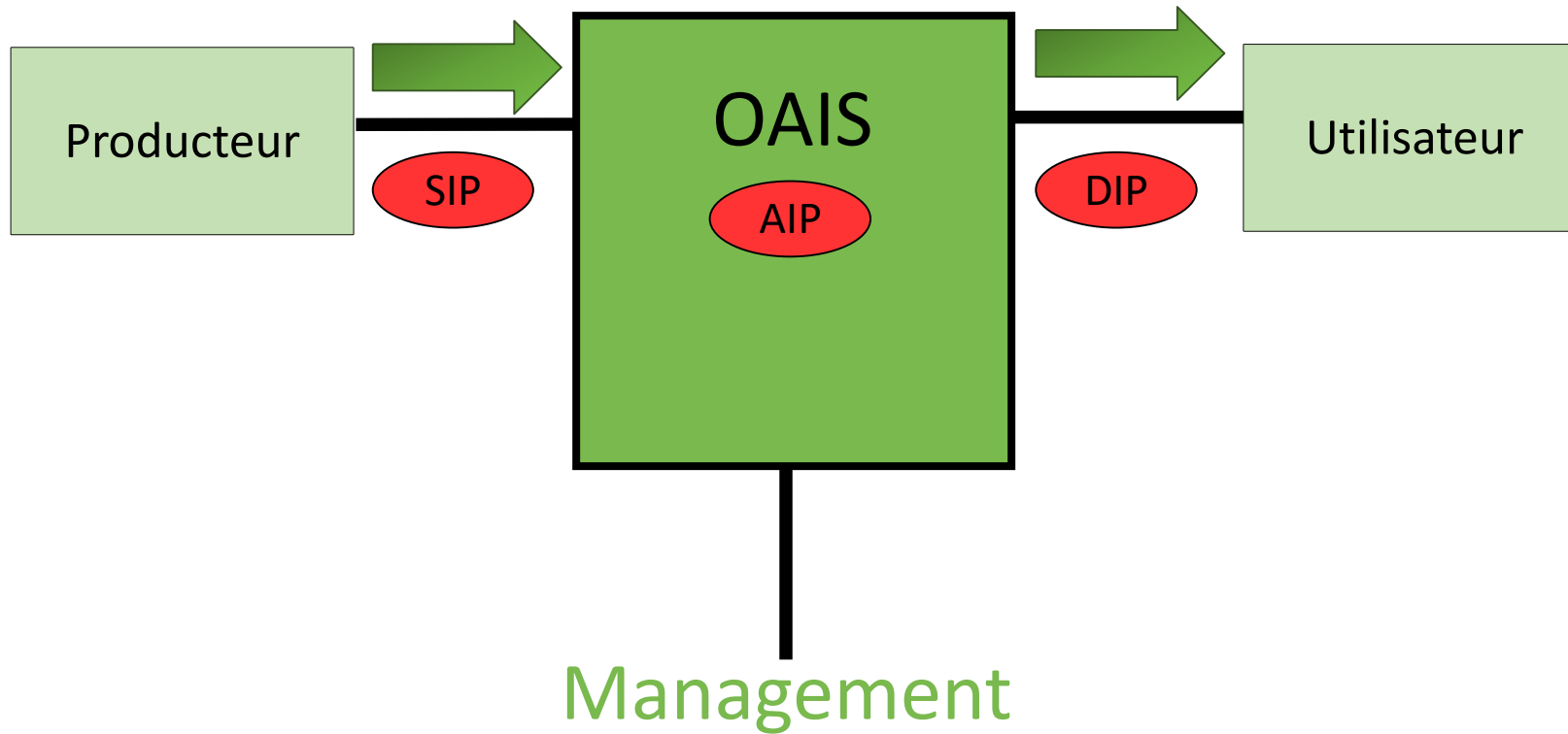
<https://public.ccsds.org/Pubs/650x0m2.pdf>
(Version 2, 2012)

[Draft Octobre 2020](#)

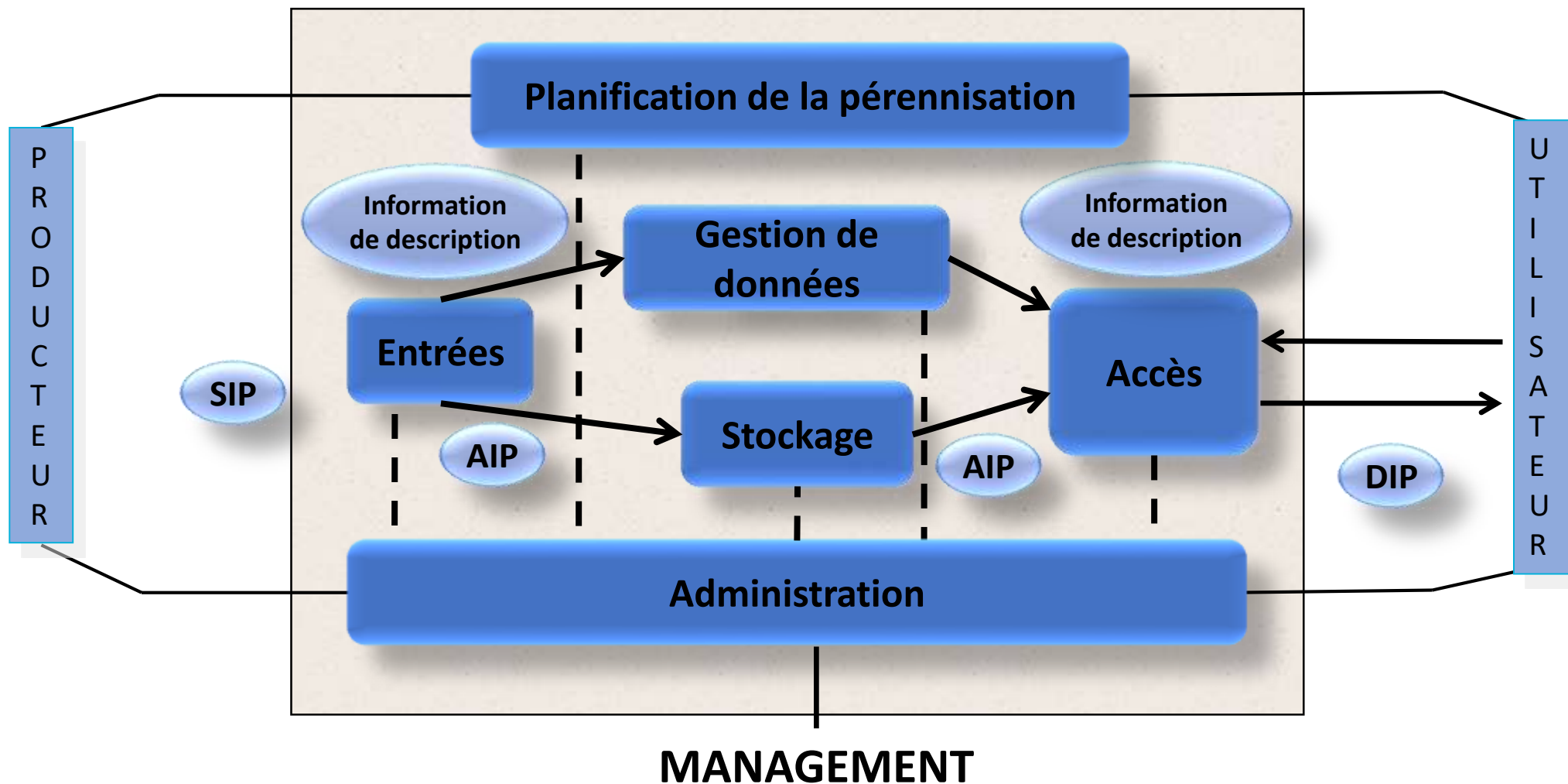
🌐 Site en français

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

L'environnement d'une archive OAIS

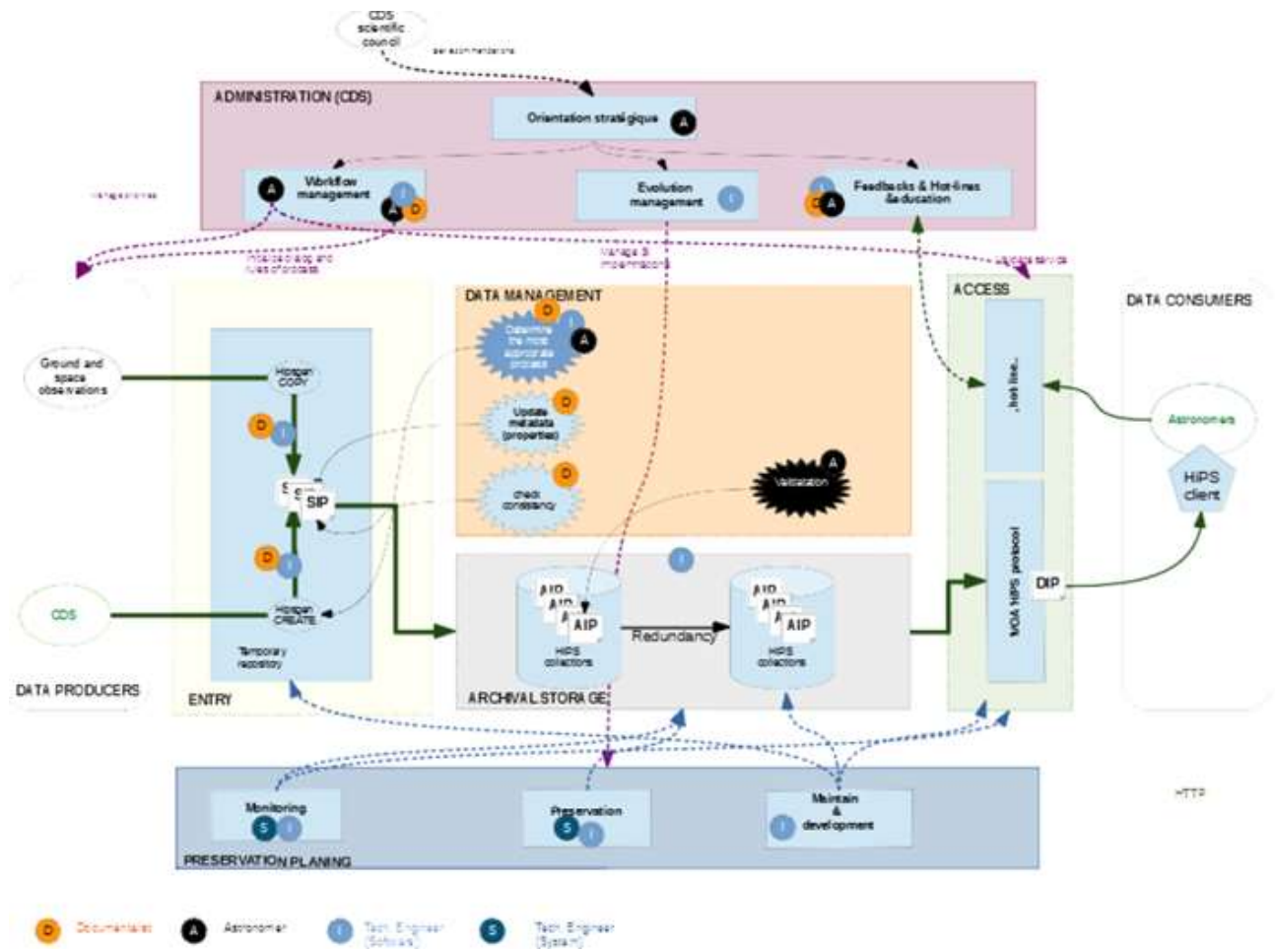


Les entités fonctionnelles de l'OAIS

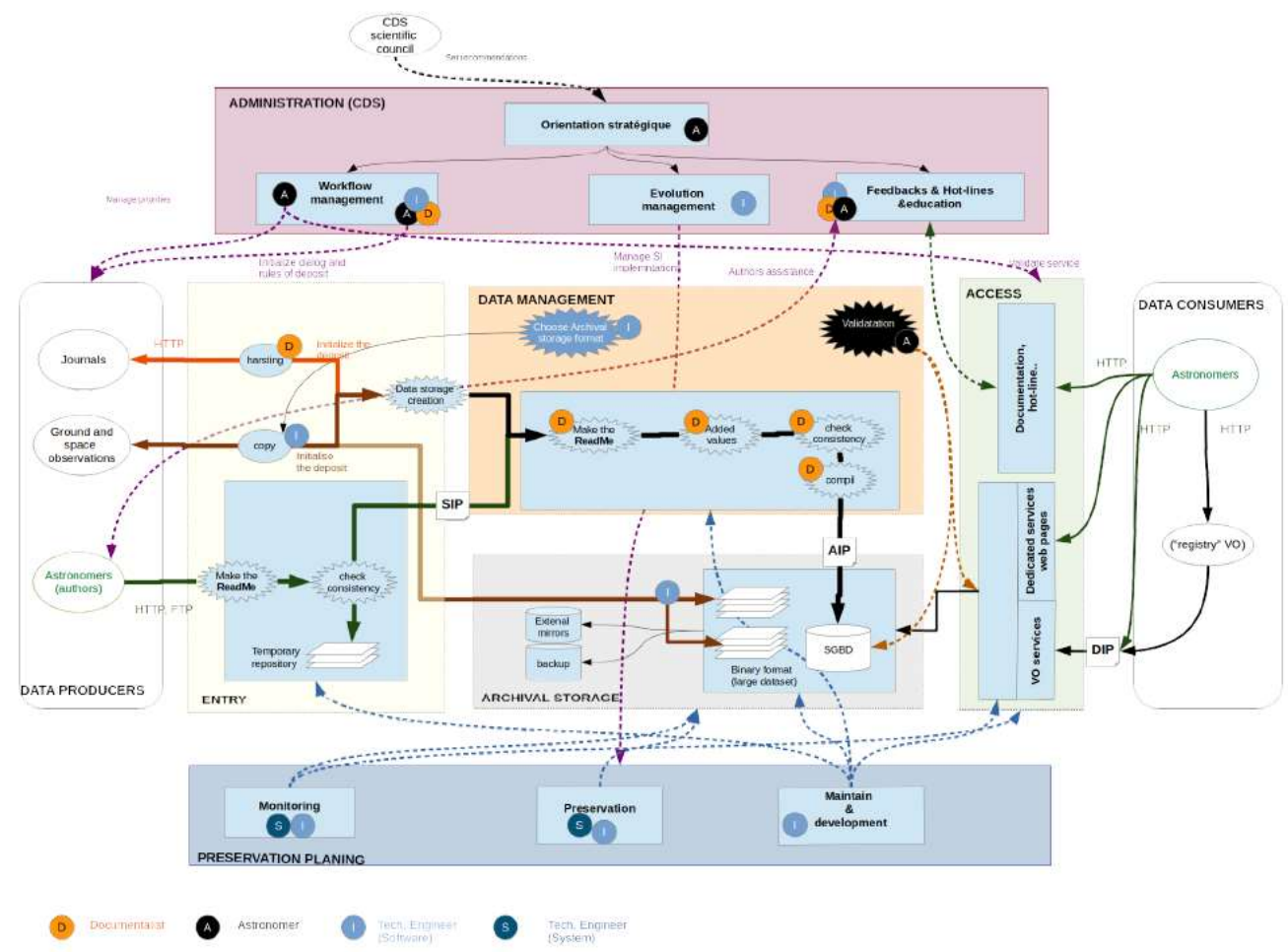




Le pipeline de données pour Aladin/CDS



Le pipeline de données pour VizieR/CDS





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 dossier 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



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

A screenshot of the CoreTrustSeal website. The header includes the CoreTrustSeal logo, navigation links (Home, About, Certification, Certified Repositories, Apply, Contact), and social media icons. The main content area is titled 'Data Repositories Requirements' and contains the following text:

CoreTrustSeal Trustworthy Data Repositories Requirements

The CoreTrustSeal Trustworthy Data Repositories Requirements reflect the characteristics of trustworthy repositories. As such, all Requirements are mandatory and are equally weighted, standalone items. Although some overlap is unavoidable, duplication of evidence sought among Requirements has been kept to a minimum where possible.

We encourage repositories to explore the CoreTrustSeal Trustworthy Data Repositories Requirements and [seek certification](#).

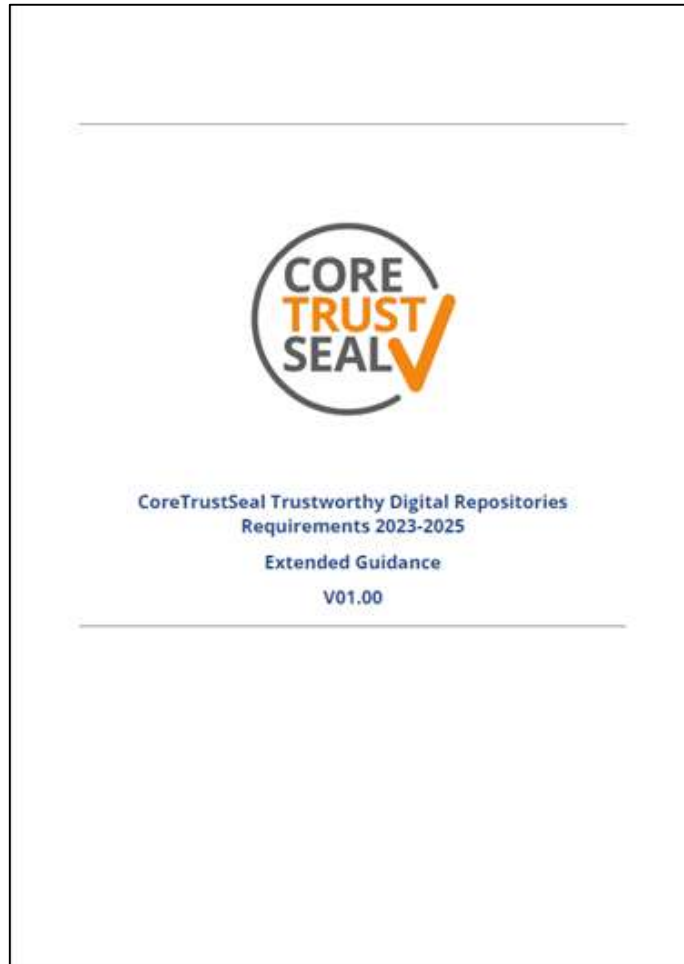
CoreTrustSeal Requirements 2023-2025:

- [CoreTrustSeal Requirements 2023-2025](#)

The following documents are available in support of the requirements:

- [CoreTrustSeal Extended Guidance 2023-2025](#) CoreTrustSeal Extended Guidance is available to facilitate the work of CoreTrustSeal reviewers and also provide more guidance to repositories undergoing a certification

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



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

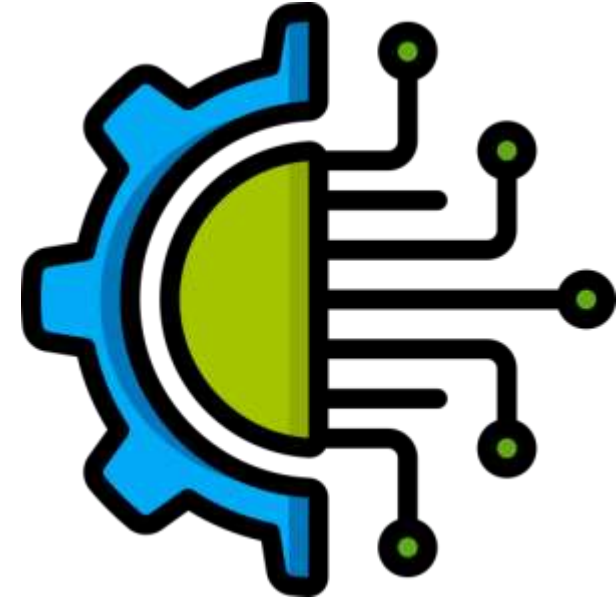
- 🌐 R7 – Provenance et authenticité des données
- 🌐 R8 – Appréciation et sélection des données
- 🌐 R9 – Assurance qualité
- 🌐 R10 – Plan de préservation
- 🌐 R11 – Processus de traitement (Workflows)
- 🌐 R12 – Découverte et identification des données
- 🌐 R13 – Réutilisation des données





Technologie

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

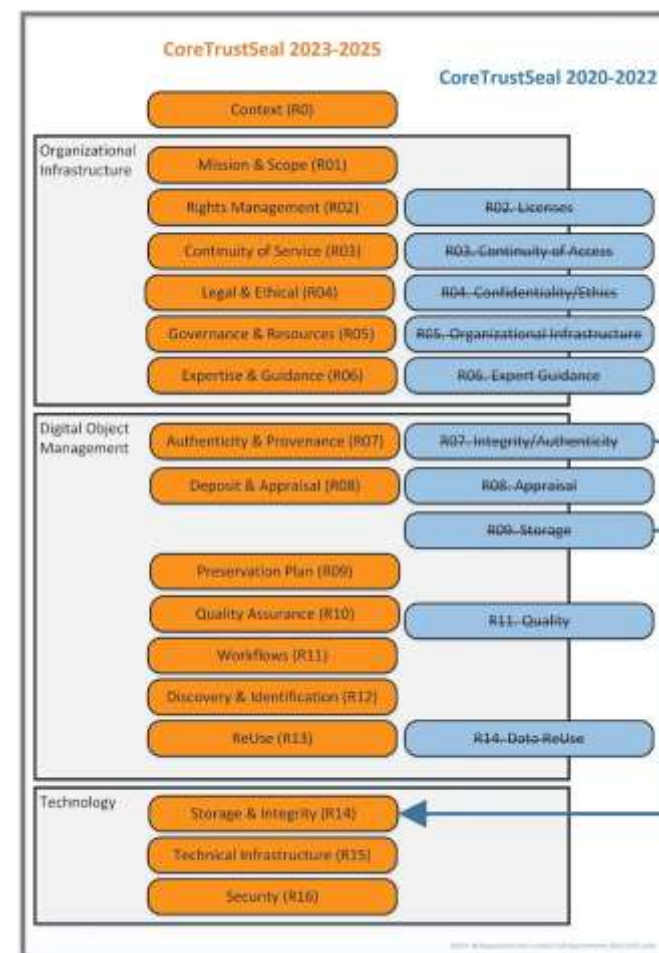


Les changements

🌐 Pas de révolution!

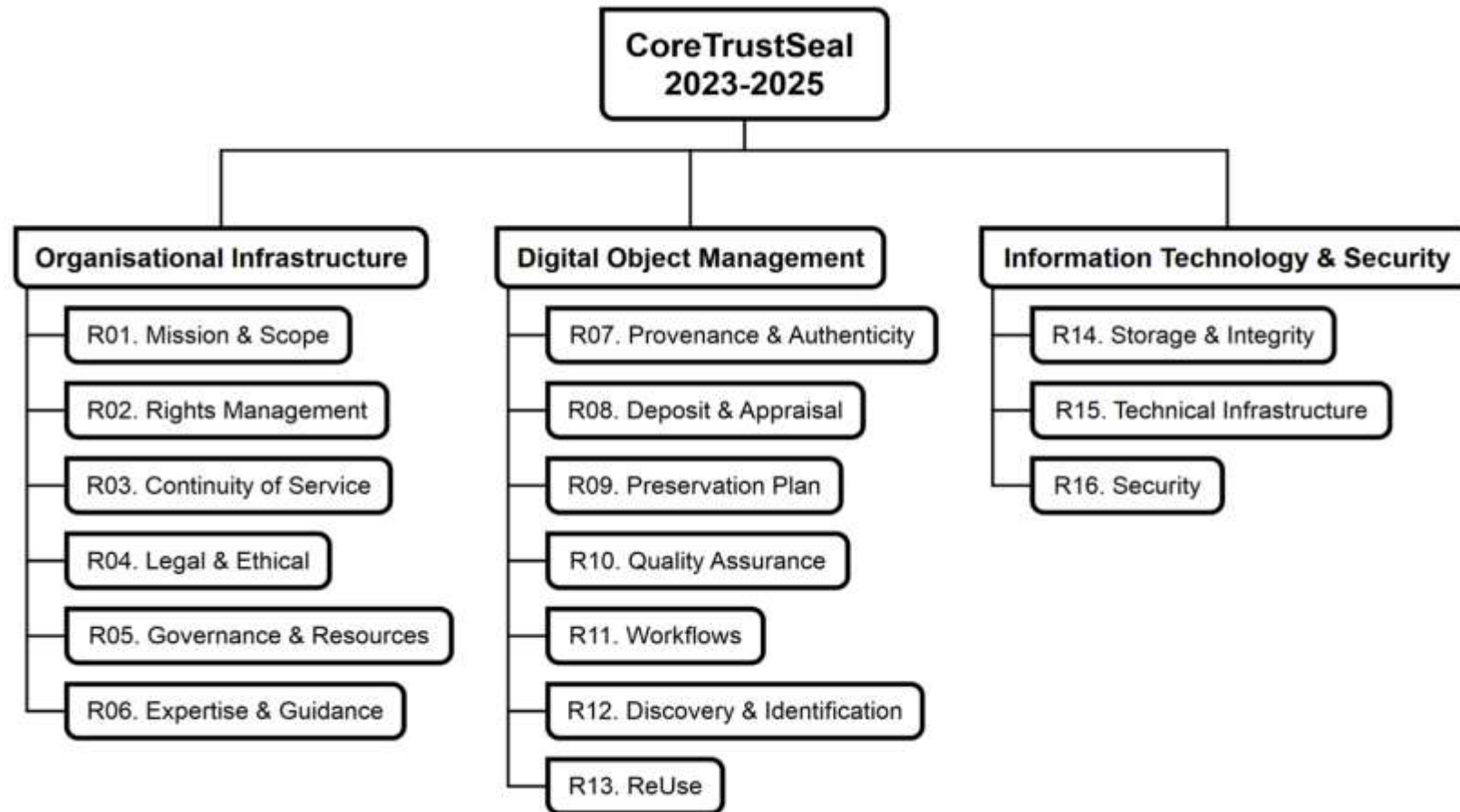
🌐 Modifications

- Du texte
- Des critères
- Deux niveaux de conformité
 - In Progress
 - Implemented



Credits: Mari Kleemola & Hervé L'Hours

Les nouveaux critères 2023-2025





Parmi les changements (1)

🌐 Une évolution du contenu du critère R0

🌐 R2 Licences >> **R02 Gestion des Droits**

The repository maintains all applicable licenses covering data access and use and monitors compliance

>> **The repository maintains all applicable rights and monitors compliance.**

🌐 R3 Continuité d'accès >> **R03 Continuité de service**

The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

>> **The Repository has a plan to ensure ongoing access to and preservation of its data and metadata.**

Parmi les changements (2)

R4 Confidentialité/Ethique > **R04 Aspects Légaux et Ethiques**

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

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

R10 Qualité des données > **R10 Assurance Qualité**

The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality related evaluations.

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



Parmi les changements (3)

R11 Workflows > **R11 Workflows**

Archiving takes place according to defined workflows from ingest to dissemination.

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

R9 Procédures de stockages documentées +dans R7 Intégrité> **R14 Stockage & Intégrité**

The repository applies documented processes and procedures in managing archival storage of the data.

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



Conclusions



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



Le rôle de RDA France

- La certification est une priorité
- 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 »
 - Première réunion le 13 octobre 2023, pendant les Journées RDA France 2023
- Liste de diffusion
<https://listes.services.cnrs.fr/wws/subscribe/rda-france-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