COPiLOtE

Un projet ANR vers la certification et la FAIRisation des Centres de Données et de Services du pôle Océan ODATIS

Conférence merIGéo

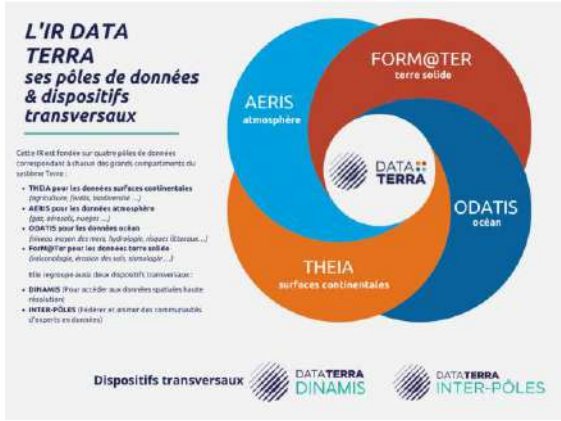
Michèle Fichaut (IFREMER) et l'équipe projet COPiLOtE

contact@odatis-ocean.fr | www.odatis-ocean.fr

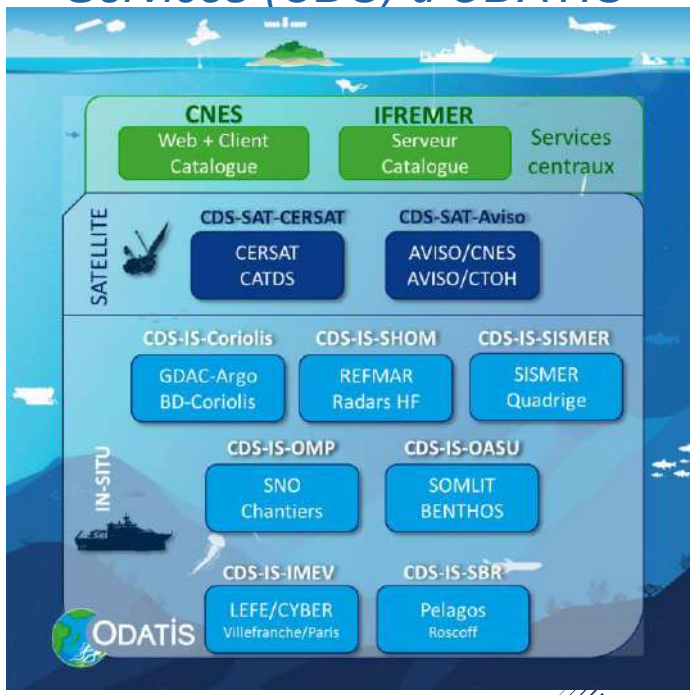
Le pôle Océan ODATIS

Missions

- Promouvoir et faciliter l'accès aux données d'observations réalisées dans l'océan ou à son interface avec les autres milieux, à partir de mesures in situ et de télédétection
- Fédérer au niveau national des activités de gestion de données et d'expertise scientifique en océanographie
- **Partenaires**



Les Centres de Données et Services (CDS) d'ODATIS



www.odatis-ocean.fr

Objectifs du projet ANR Flash COPiLOtE (2020-2022)

- COPiLOtE : **C**ertificati**O**n **P**ô**L**e **O**c**E**an

- Accompagner les Centre de Données et de Services (CDS) ODATIS à déposer un dossier de certification auprès de **Core Trust Seal (CTS)**

- Réaliser une évaluation du caractère **FAIR** des données gérées par les CDS ODATIS

Copilote – certification *Core Trust Seal*

- CTS est un organisme de certification des entrepôts de données, né de la fusion de 2 systèmes de certifications (DSA & WDS), sous l'égide de la RDA

2008-2018

2011-2018

https://www.coretrustseal.org/

Critères pour être certifié

L’entrepôt de données est évalué selon **16 critères** différents organisé en 3 thèmes principaux:

- Organisation de l’infrastructure
- Gestion des données numériques (données et métadonnées)
- Technologie et sécurité de l’infrastructure

R1	Mission/champ d'application	R9	Procédures de stockage documentées
R2	Licences	R10	Plan de préservation
R3	Continuité d'accès	R11	Qualié de la données
R4	Confidentialité/Ethique	R12	Wokflow de l'ingestion à la dissémination
R5	Infrastructure organisationnelle	R13	Découverte des données et identification
R6	Avis d'experts	R14	Réutilisation des données
R7	Intégrité at authenticité des données	R15	Infrastructure technique
R8	Pertinence et compréhensibilité des données	R16	Sécurité

Intérêt de la certification

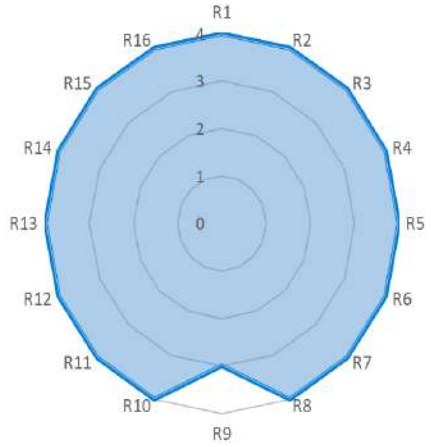
- Gage de qualité et de fiabilité
- Confiance pour les producteurs et utilisateurs de données
- Partage de bonnes pratiques
- Application des principes
- Gains en performance : interopérabilité
- Découverte des données et citation améliorées
- Pratiques et processus améliorés

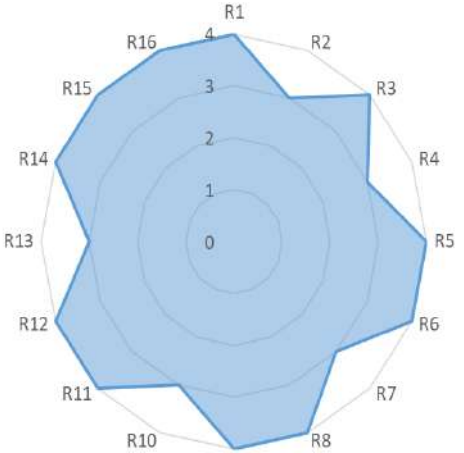
Méthode appliquée dans le cadre de COPiLOtE

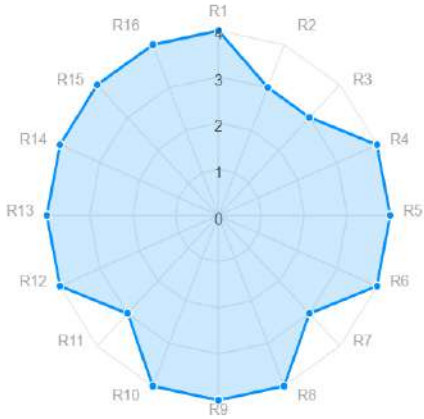
- CDS ODATIS
  - **formation RDA** sur la certification
  - **une lecture guidée** des critères CTS (cadre ODATIS)
- Accompagnement par le projet COPiLOtE pour la préparation du dossier CoreTrustSeal
  - Réunions de suivi régulières pour préparer le document (réponse critère par critère)
  - Auto-évaluation de tous les critères à la fin
  - Auto-évaluation croisée pour certains CDS

Auto évaluation des CDS

- Chaque CDS s'auto-évalue sur les 16 critères CTS et se donne **une note de 0 à 4**
  - 0 : non applicable
  - 1 : Non envisagé
  - 2 : En cours d'examen
  - 3 : En cours de mis en œuvre
  - 4 : Entièrement mis en œuvre
- Pour pouvoir prétendre à la certification il faut une note de **minimum 3** à chacun des 16 critères







COPiLOtE – Résultats certification

- 4 CDS ont déposé leur dossier de certification en octobre 2022

SOMLIT (données d'océanographie côtière) et **KIDA** (astrophysique)
 Retour revue avec commentaires

SNO **PHYTOBS** (phytoplancton) et SNO **BENTHOBS (**macrofaune)
 Retour revue avec commentaires

1 filière : REFMAR (l'observation in-situ du niveau de la mer)
 En cours de revue

1 filière AVISO+ (altimétrie de l'océan)
 En cours de revue

- 3 CDS ont fait leur demande de renouvellement:

Coriolis, SISMER & Cersat Retour revue avec commentaires, Réponse envoyée

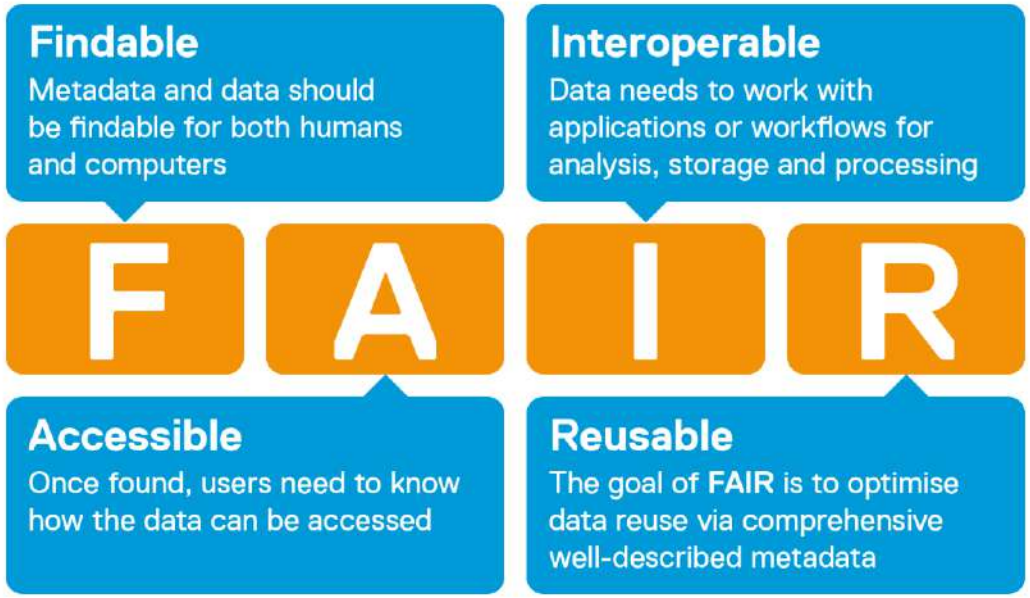
Bilan de ce travail de certification

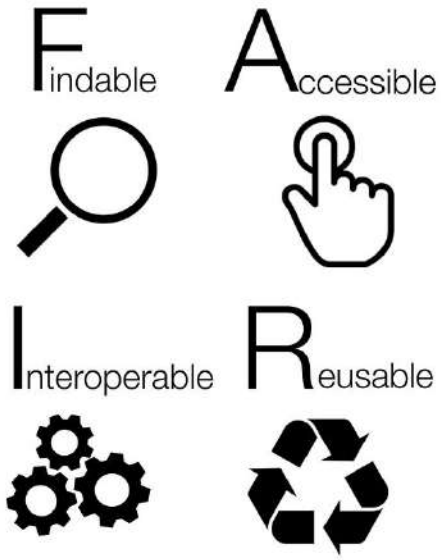
- Les quatre CDS ont constaté son apport bénéfique principalement
  - Clarification écrite des flux de données
    - Ajout d'information
    - Ajout de documents annexes utiles
    - Prise en compte de l'ensemble des processus
  - Clarification des procédures internes concernant les filières évaluées
    - Mise en place de nouvelle procédure de réplication par exemple
  - Implication et collaboration de nombreux interlocuteurs par CDS
- Difficultés rencontrées
  - Définition du périmètre de la certification, définition de la granularité des informations à décrire
  - Un peu chronophage

COPiLOtE – Auto-évaluation FAIR des CDS ODATIS

- Les principes FAIR : un ensemble de recommandations pour gérer les données de la recherche et les rendre

  Facile à trouver, **A**ccessible, **I**nteropérable et **R**éutilisable





COPiLOtE – Auto-évaluation FAIR des CDS ODATIS

- Un certain nombre d'initiatives différentes travaillent actuellement à la définition de cadres, de méthodes et de critères d'évaluation du caractère FAIR des données :
  - FAIR Implementation Profile
  - FAIRsFAIR
  - RDA - FAIR Data Maturity Model WG

COPiLOtE : Outil utilisé RDA - FAIR Data Maturity Model

- RDA - FAIR Data Maturity Model WG
  - RDA FAIR Data Maturity Model Specification and Guidelines Recommendation : https://doi.org/10.15497/rda00050
- Modèle d'auto-évaluation générique pour mesurer le niveau de maturité d'un jeu de donnée
- Approche proposée par la RDA/FDMM la plus adaptée au contexte ODATIS
- Utilisée pour l'auto-évaluation FAIR des CDS ODATIS dans le cadre du projet COPiLOtE

Indicateurs du FAIR DATA MATURITY MODEL (FDMM)

- 41 indicateurs au total
  - **F** : 7 indicateurs
  - **A** : 12 indicateurs
  - **I** : 12 indicateurs
  - **R** : 10 indicateurs

Les indicateurs sont classifiés comme:
- Essentiel: de la plus haute importance
- Important: Accroit substantiellement le caractère FAIR
- Utile: bien à avoir, mais pas indispensable

	Principle				
Priority	Findable	Accessible	Interoperable	Reusable	Grand Total
Essential	7	8	0	5	20
Important	0	3	7	4	14
Useful	0	1	5	1	7
Grand Total	7	**12**	**12**	**10**	**41**

Indicateurs du FAIR DATA MATURITY MODEL (FDMM)

Facile à trouver : 7 indicateurs, tous essentiels

Identifiants

1. **F1-01M**: Les métadonnées sont identifiées par un identifiant pérenne
2. **F1-O1D**: Les données sont identifiées par un identifiant pérenne
3. **F1-O2M**: Les métadonnées sont identifiées par un identifiant mondialement unique pérenne
4. **F1-O2D**: Les données sont identifiées par un identifiant mondialement unique pérenne
5. **F2-O1M**: Des métadonnées riches permettent la découverte des données
6. **F3-O1M**: Les métadonnées comprennent l'identifiant des données
7. **F4-O1M**: Les métadonnées sont proposées de manière à pouvoir être moissonnées et indexées

Indicateurs du FAIR DATA MATURITY MODEL (FDMM)

12 indicators for ACCESSIBLE (8 essentiels, 3 importants, 1 utile)

Protocole d'accès, Authetntification, Autorisation, identifiants

8. **A1-01M**: Les métadonnées contiennent des informations permettant à l'utilisateur d'accéder aux données
9. **A1-02M**: Les métadonnées sont accessibles manuellement (i.e. avec une intervention humaine)
10. **A1-02D** : Les données sont accessibles manuellement (i.e. avec une intervention humaine)
11. **A1-03M**: L'identifiant de métadonnées renvoie à un enregistrement de métadonnées
12. **A1-03D**: L'identifiant des données renvoie à un objet numérique
13. **A1-04M**: Les métadonnées sont accessibles via un protocole standardisé (e.g. HTTP, FTP, …)
14. **A1-04D**: Les données sont accessibles via un protocole standardisé (e.g. HTTP, FTP, …)
15. **A1-05D**: Les données sont accessibles automatiquement (i.e. par un programme informatique)
16. **A1.1-01M**: Les métadonnées sont accessibles via un protocole d'accès libre
17. **A1.1-01D**: Les données sont accessibles via un protocole d'accès libre
18. **A1.2-01D**: Les données sont accessibles via un protocole d'accès qui prend en charge l'authentification et l'autorisation
19. **A2-01M**: Il est garanti que les métadonnées restent disponibles après que les données ne le soient plus

Indicateurs du FAIR DATA MATURITY MODEL (FDMM)

12 indicateurs for Interopérable (7 importants, 5 utiles)

Standard, FAIR, Lisible par machine, linked data

20. **I1-01M**: Les métadonnées utilisent une représentation des connaissances exprimée dans un format standardisé
21. **I1-01D**: Les données utilisent une représentation des connaissances exprimée dans un format standardisé
22. **I1-02M**: Les métadonnées utilisent une représentation des connaissances compréhensible par une machine
23. **I1-02D**: Les données utilisent une représentation des connaissances compréhensible par une machine
24. **I2-01M**: Les métadonnées utilisent des vocabulaires conformes aux principes FAIR
25. **I2-01D**: Les données utilisent des vocabulaires conformes aux principes FAIR
26. **I3-01M**: Les métadonnées incluent des références à d'autres métadonnées
27. **I3-01D**: Les données incluent des références à d'autres données
28. **I3-02M**: Les métadonnées incluent des références à d'autres données
29. **I3-02D:** Les données incluent des références qualifiées à d'autres données
30. **I3-03M**: Les métadonnées incluent des références qualifiées à d'autres métadonnées
31. **I3-04M**: Les métadonnées incluent des références qualifiées à d'autres données

Indicateurs du FAIR DATA MATURITY MODEL (FDMM)

10 indicateurs for Réutilisable (5 essentiels, 4 importants, 1 utile)

Licence, norme communautaire

32. **R1-01M** : Une pluralité d'attributs précis et pertinents sont fournis pour permettre la réutilisation
33. **R1.1-01M**: Les métadonnées comprennent des informations sur la licence sous laquelle les données peuvent être réutilisées
34. **R1.1-02M**: Les métadonnées font référence à une licence de réutilisation standard
35. **R1.1-03M**: Les métadonnées font référence à une licence de réutilisation compréhensible par une machine
36. **R1.2-01M**: Les métadonnées comprennent des informations sur la provenance selon des normes spécifiques à la communauté
37. **R1.2-02M**: Les métadonnées incluent des informations de provenance selon un langage intercommunautaire
38. **R1.3-01M**: Les métadonnées sont conformes à une norme communautaire
39. **R1.3-01D**: Les données sont conformes à une norme communautaire
40. **R1.3-02M**: Les métadonnées sont exprimées conformément à une norme communautaire compréhensible par les machines
41. **R1.3-02D**: Les données sont exprimées conformément à une norme communautaire compréhensible par les machines

2 méthodes d’évaluation du FDMM

- Mesurer la réussite ou l’échec: Déterminer si une ressource évaluée répond aux exigences d'un indicateur exprimé sur une échelle binaire de réussite ou d'échec.

- Mesurer la maturité : Fournir une mesure dans laquelle une ressource évaluée répond aux exigences d'un indicateur exprimé suivant l'échelle
  - 0 = non applicable,
  - 1 = non envisagé,
  - 2 = en cours d'examen,
  - 3 = en cours de mise en œuvre,
  - 4 = entièrement mis en œuvre.

→ Même échelle de mesure que la certification Core Trust Seal

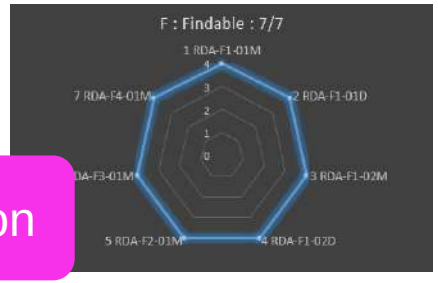
COPiLOtE : Auto-évaluation FAIR des CDS ODATIS

- Questionnaire d'auto-évaluation avec chaque filière – Réunions pour répondre au questionnaires (2 réunions par filière)
  - **CDS-IS-SBR** : pour PHYTOBS et BENTHOBS,
  - **CDS-IS-CORIOLIS** : pour ARGO, DBCP, OceanSITES-PIRATA, GOSUD-Ferry Box, Gliders
  - **CDS-IS-IMEV** : pour Lefe-Cyber
  - **CDS-IS-OASU** : pour SOMLIT et KIDA
  - **CDS-IS-SISMER** : pour les filières SISMER : Données géographiques (Sextant), Données des campagnes : de physique chimie et Géosciences, Catalogue des campagnes, Quadrige/SURVAL (monitoring côtier), Echantillons biologiques et géologiques
  - **CDS-IS-Shom** : pour RONIM/REFMAR
  - **CDS-SAT-CERSAT** : pour l'ensemble des produits + CATDS
  - **CDS-SAT-AVISO** : pour AVISO+

Auto-évaluations – exemple du PRINCIPE F

ARGO

DBCP



LEFE CYBER

SOMLIT

KIDA

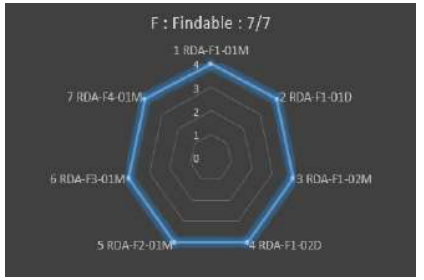
En cours de labélisation

DOI ont été prégénérés mais ne pointent pas vers des Landing Pages

PHYTOBS

BENTHOBS

REFMAR

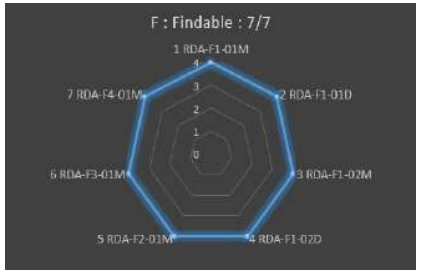


DOI non disponible pour le moment, pas de moissonnage possible

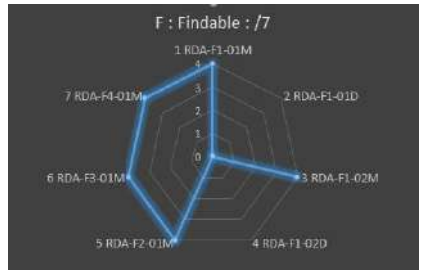
CERS

TILLONS

GEOSCIENCES



DONNEES GEO



SURVAL

PHYSIQUE-CHIMIE

13/17

Intérêt de cette auto-Evaluation

- Pour les CDS
  - Exercice très intéressant : recul sur ses pratiques
  - Permet d’identifier les améliorations à apporter aux ensembles de données, en particulier concernant les critères essentiels qui sont "obligatoires" pour avoir des données FAIR.
- Pour le pôle ODATIS
  - Vision globale sur le caractère FAIR des données de l’ensemble des CDS
  - Identification d’outil à mettre en œuvre pour améliorer ce caractère FAIR

En conclusion de cette étude

- Le caractère FAIR des données dépend énormément de l'outil de diffusion
- La Participation à des projets européeens (SeaDataNet, ENVRI FAIR, …) améliore souvent les pratiques
- Pistes d'amélioration
  - Utiliser des vocabulaire FAIR (pas exemple en utilisant des vocabs existants : NERC-BODC, GCMD…)
  - Utiliser les principes de données liées, SPARQL endpoint, web sémantique
  - Enrichir les métadonnées avec références qualifiées à d'autres métadonnées et données quand c'est possible (ORCID, Archimer, EDMO …)
  - Enrichir les métadonnées avec des liens vers d'autres outils de diffusion (SeaDataNet, EMODnet, Sextant…)

Merci de votre attention

contact@odatis-ocean.fr | www.odatis-ocean.fr