

École Thématique DATA SDUE 2022
Guide de Survie dans la jungle des données en Sciences de l'Univers et de l'Environnement (SDUE) :

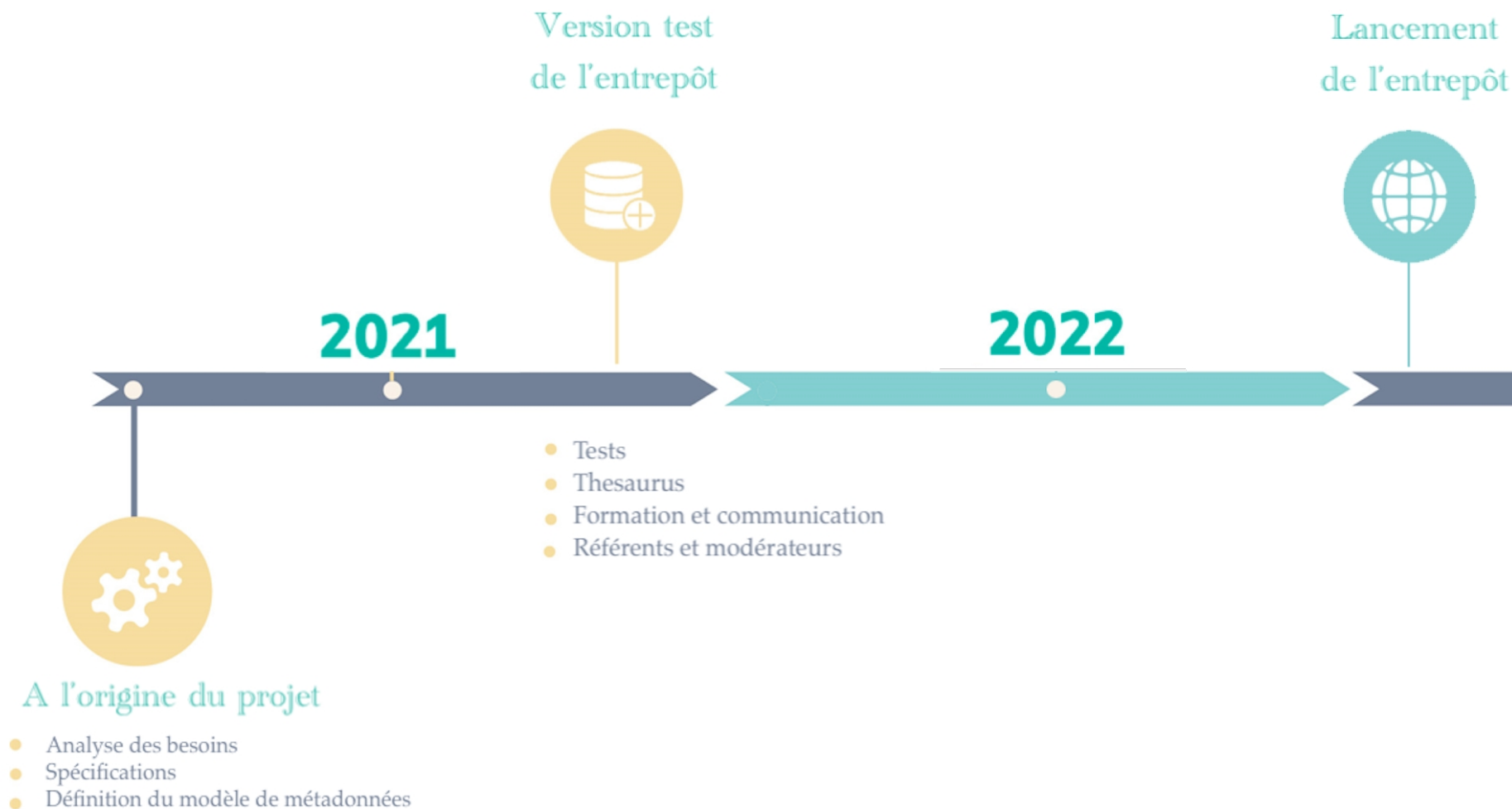
Comment gérer les données pour les valoriser?

Session « Choix de l'entrepôt de données pour dépôt et diffusion des données FAIR »
Exemple de la construction de l'entrepôt de données de longue traîne: DataTerra

C. Pierkot - H. Bressan

Équipe projet

- Véronique Bertrand – RESIF
- Hélène Bressan – BRGM
- Christelle Pierkot - Data Terra
- Marine Vernet – Data Terra



La genèse de l'entrepôt

Besoins des pôles

=> Gestion de données produites dans des entités qui ne disposent pas de moyens pour les FAIRiser et les stocker.

=> Produits à valeur ajoutée issus de données d'observation produites par les pôles, en vue de publication et l'obtention d'un DOI.

=> Données issues de projets nationaux et internationaux dont Data Terra est porteur.

Enquête communautés

⇒ Données :

- ⇒ Volumétrie : Fichier : entre 4Mo et 1To et Dépôts : entre 20 Mo et 10 Po
- ⇒ Formats hétérogènes

⇒ Droits d'accès :

- ⇒ Ouvert 76 %
- ⇒ Embargo 50 % avec durée variable en moyenne 2 ans
- ⇒ En demande d'accès : 45 %

⇒ Visibilité et accessibilité :

- ⇒ Moissonnage des entrepôts existants
- ⇒ Référencement dans des systèmes nationaux et/ou internationaux

⇒ Divers

- ⇒ Facilité d'utilisation et simplicité du dépôt
- ⇒ Gestion des versions
- ⇒ Exposition des métadonnées selon différents formats
- ⇒ Statistique de téléchargement et d'utilisation des dépôts

Analyse des solutions existantes

⇒ Retours d'expérience

- ⇒ Seanoe
- ⇒ DataSud

⇒ Analyse de l'existant

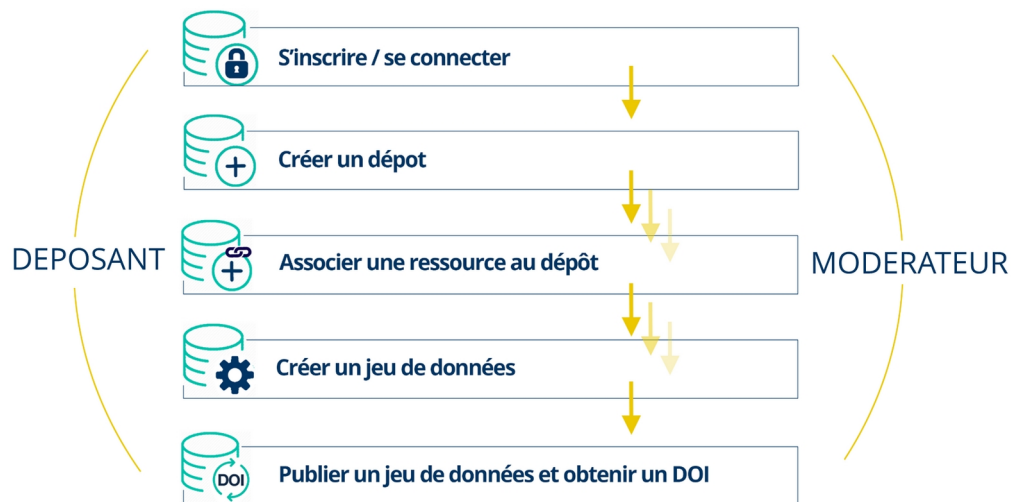
- ⇒ Pangaea
- ⇒ Zenodo
- ⇒ Dataverse CIRAD

Fonctionnalités

Spécifications

- ⇒ Données géoréférencées
- ⇒ Métadonnées normalisées (ISO 19115)
- ⇒ Gestion multilingue
- ⇒ DOI et Citation
- ⇒ Identification des auteurs
- ⇒ Niveaux de granularité
- ⇒ Vocabulaires
- ⇒ Licence CC BY par défaut
- ⇒ Volumétrie (5Go par fichier, 100Go par dépôt)
- ⇒ Contraintes d'accès et embargo
- ⇒ Formats de données non contraint
- ⇒ Authentification Renater
- ⇒ Recherche à facettes
- ⇒ Associer des ressources aux données

Workflow de dépôt



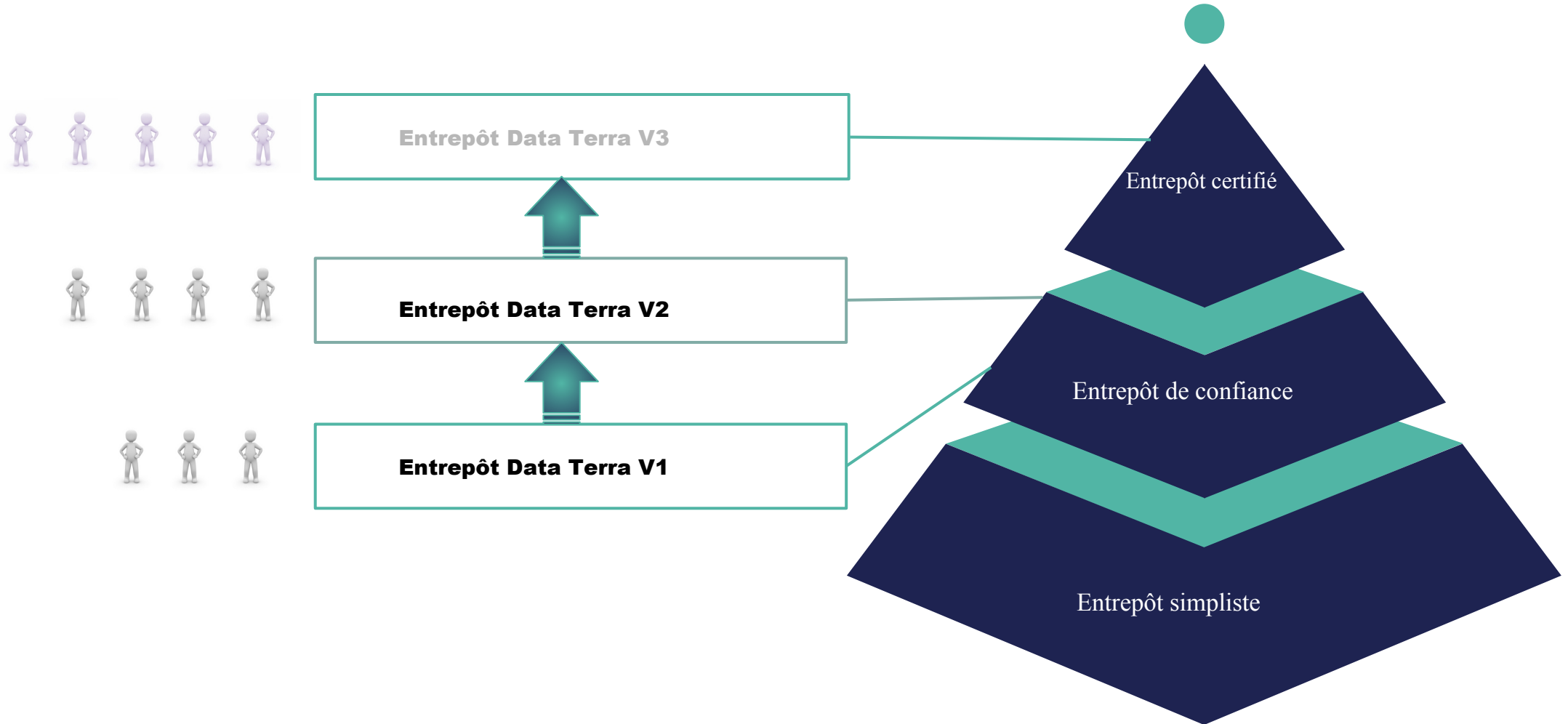
Améliorations prévues

- ⇒ Connexion software Heritage – Hal
- ⇒ Gestion des versions
- ⇒ DataPaper
- ⇒ Analyse antivirus
- ⇒ Statistique d'utilisation
- ⇒ Prise en compte retours utilisateurs

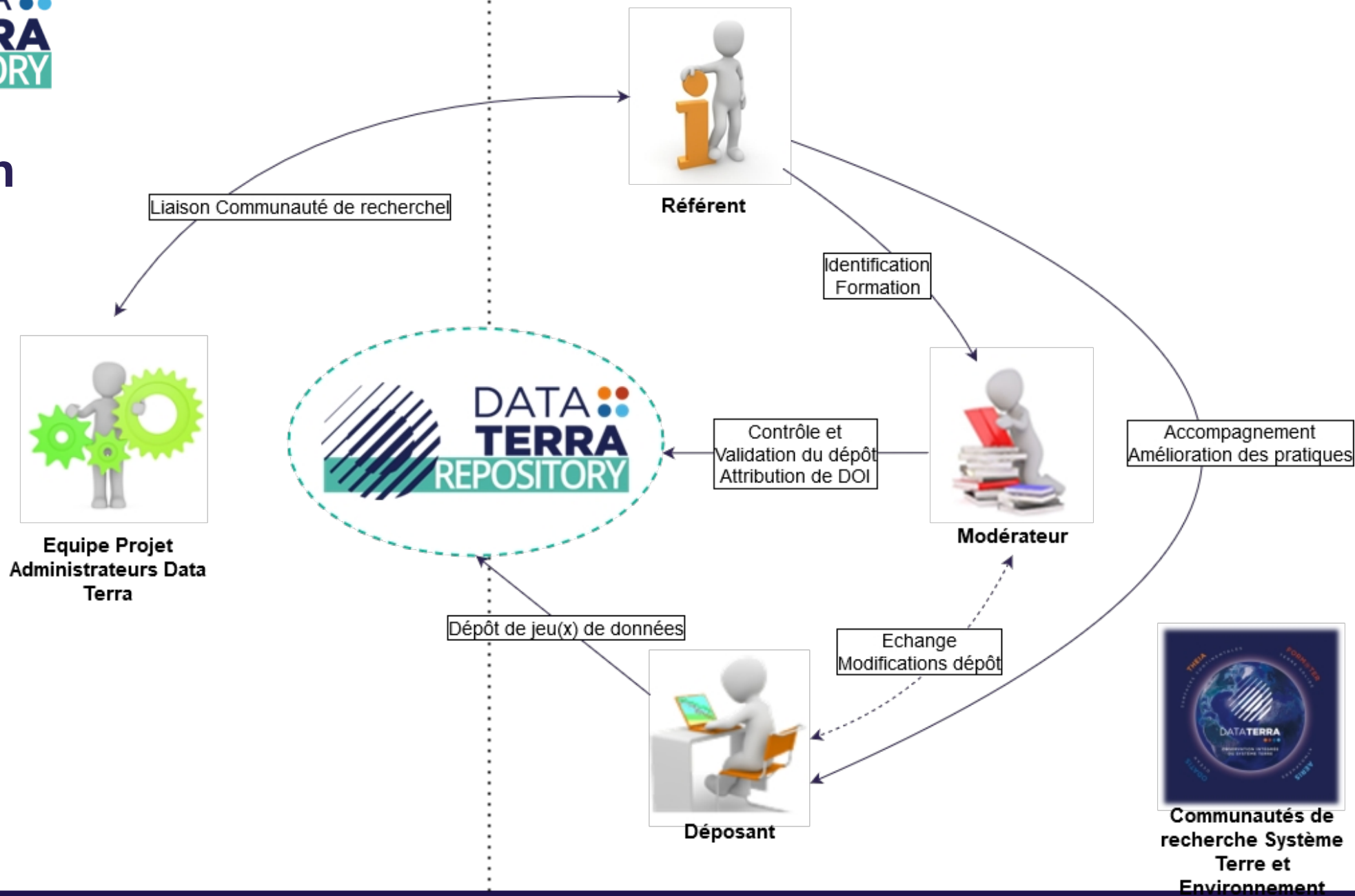


En route vers la certification

Positionnement entrepôt DataTerra



Organisation



Accompagnement / Communication

Support

- ⇒ Guide de dépôt
 - ⇒ Guide de bonnes pratiques
 - ⇒ Guide de modération
 - ⇒ Qualité attendue
 - ⇒ Périmètre
- Communautés

Promotion

- ⇒ Page web : <https://www.data-terra.org/data-repository/>
- ⇒ Flyer

Ateliers

- ⇒ Ateliers Resif et GNSS
- ⇒ Université de Lille 12/05 – Fabrique de la Science Ouverte Session Entrepôts de données disciplinaires
- ⇒ Ateliers de la données



Genèse

- Besoins identifiés pour entreposer les données de longue traîne
- Intégration de V. Bertrand (Résif) dans le GT entrepôt Data Terra
- Présentation du projet d'entrepôt de données Data Terra lors des Journées Scientifiques et Techniques Résif, novembre 2021.

Méthode

- Recherche de volontaires
- Choix de jeux de données ciblés
- Formation aux différentes rôles
- Beta-test de dépôt réalisé de bout en bout.

Retours

- Très instructif pour améliorer l'outil !
- Prise de conscience de l'intérêt de l'ingénierie de la donnée :
 - Identification des standards de description (méta)données
 - Format des (méta)données
 - Vocabulaire contrôlé, thesaurus, ontologie...
- Des problèmes de périmètre : diversité de la longue traîne

Tests avec les autres communautés

=> ODATIS / SEANOE

- ⇒ Moissonnage - Compatibilité des schémas ?
- ⇒ Retour d'expérience en terme d'organisation à considérer pour les autres pôles.

=> THEIA ET AERIS

- ⇒ Définition du périmètre
- ⇒ Mise en place de l'organisation
- ⇒ Vocabulaires : Mots clés et Thématiques

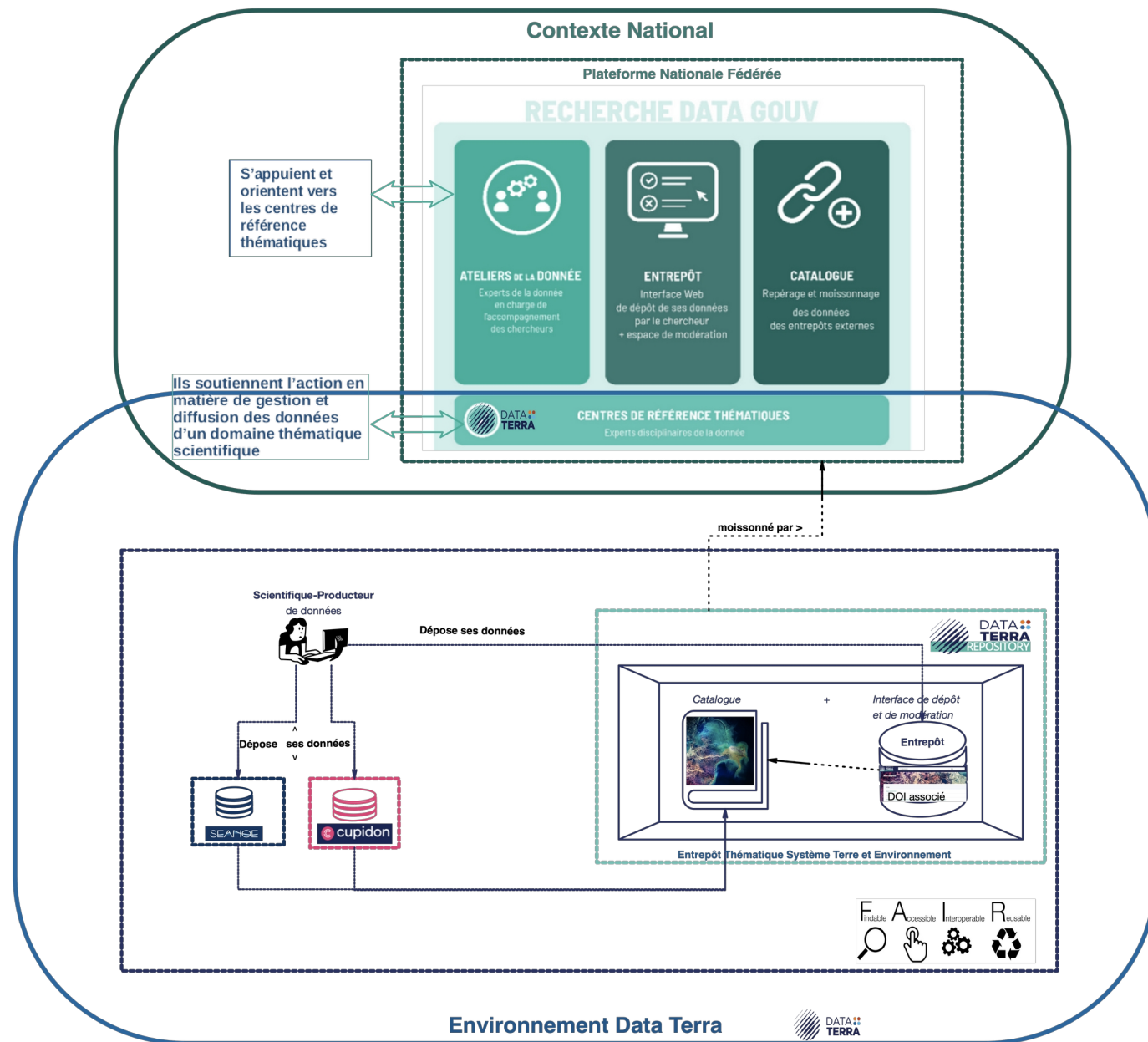
=> CLIMERI

- ⇒ Communauté structurée et internationale
- ⇒ Groupe de travail interne
- ⇒ Volumétrie importante

=> PNDB

- ⇒ Compatibilité des niveaux de granularité ?
- ⇒ Moissonnage d'entrepôts existants ?

Entrepôt de données DataTerra dans l'écosystème de la Science Ouverte



Merci de votre attention

Pour nous contacter : entrepot@data-terra.org

Liste de diffusion :

https://groupes.renater.fr/sympa/subscribe/entrepot_dataterra

Site web : <https://www.data-terra.org/donnees-services/entrepot-de-donnees-data-terra/>



Entrepôt thématique
des données du système Terre et Environnement

Données géoréférencées

Métadonnées normalisées

Science ouverte

DOI

FAIR

Visibilité

Citation

Réutilisation

**DATA
TERRA
REPOSITORY**

Pour qui ?

- Les scientifiques du système Terre et Environnement
- Les partenaires nationaux, européens et internationaux

Quels types de données ?

Données géoréférencées :

- Issues de recherches en vue de publication
- Issues de projets

Un processus en 4 étapes :

- 1 Créer un dépôt
- 2 Associer les jeux de données
- 3 Associer des ressources
- 4 Attribuer un DOI

→ Accompagnement par les équipes ressources

www.data-terra.org/dataterra-repository/

Ne pas jeter sur la voie publique.