Observatoire Midi-Pyrénées

LES DOI

IDENTIFIANTS PÉRENNES ET CITATION DE DONNÉES

Guillaume Brissebrat, responsable du SEDOO

6 juillet 2018 – Séminaire OVGSO - IRAP

DOI – Digital Object Identifier

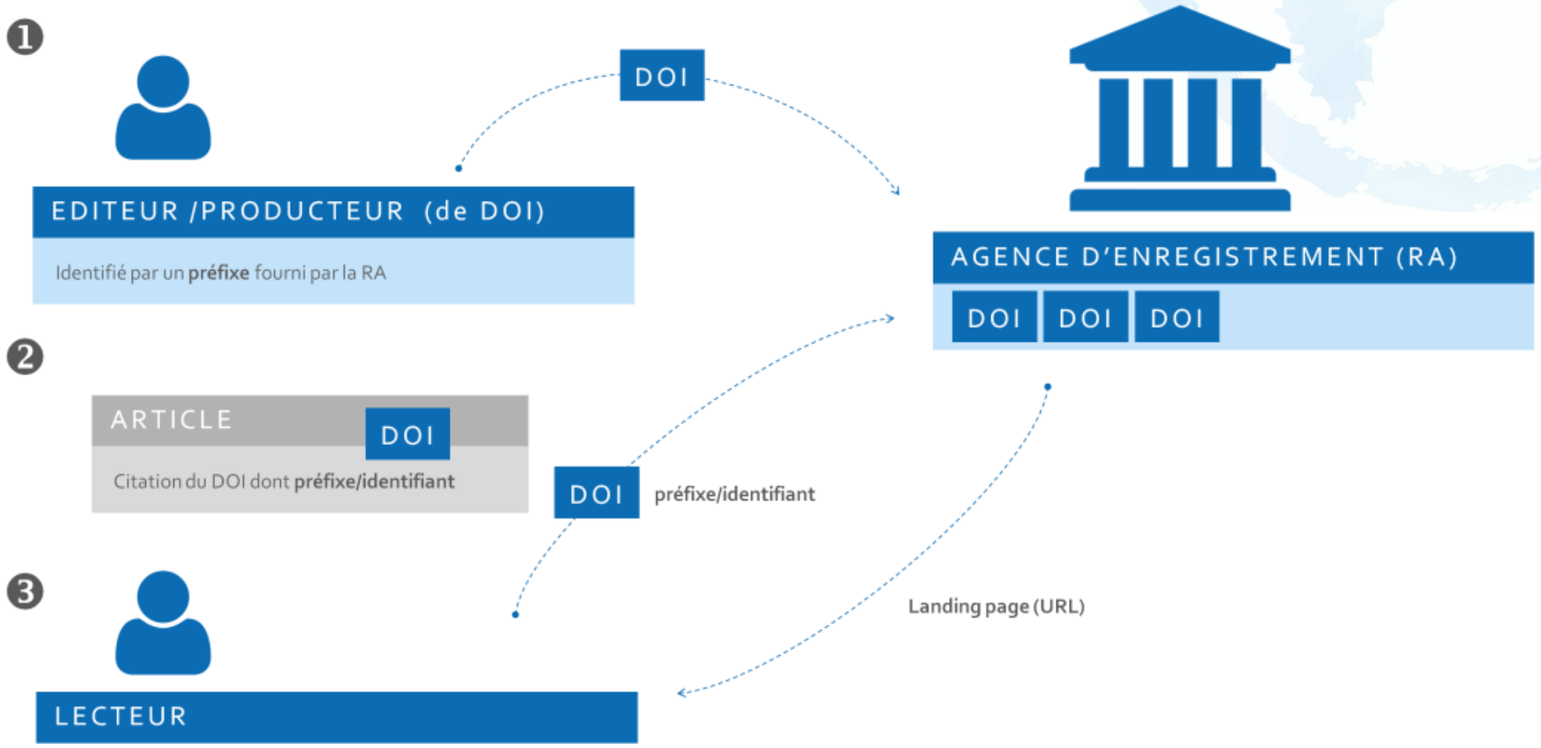
- Identifiant unique attribué de manière pérenne à un *objet digital*
  - Publication, **jeu de données**, rapport, documentation, etc.
- Objectifs
  - Accéder à la ressource sur le long terme
  - Faciliter la découverte, le partage, la réutilisation des données
  - Faciliter la citation

Non pérennité des URL => nécessité d'un identifiant pérenne

- Contrôlé et géré par IDF (International DOI Foundation)
  - 10 agences d'enregistrement dont DataCite qui fournit les solutions les mieux adaptées pour les jeux de données scientifiques
- Standard international (ISO 26324:2012)

Fonctionnement

- Les 3 étapes de l'utilisation d'un DOI :
  1. Création et enregistrement dans une RA
  2. Citation dans un article
  3. Résolution - via la RA



- Une fois créé, les informations associées au DOI peuvent être modifiées, notamment l'URL de sa landing page

L'agence d'enregistrement DataCite

- Consortium international à but non lucratif
- Créé officiellement le 1er décembre 2009 à Londres
- Centré sur la valorisation des données de recherche
  - Faciliter l'accès
  - Faciliter leur citation
  - Promouvoir leur publication
  - Soutenir leur archivage

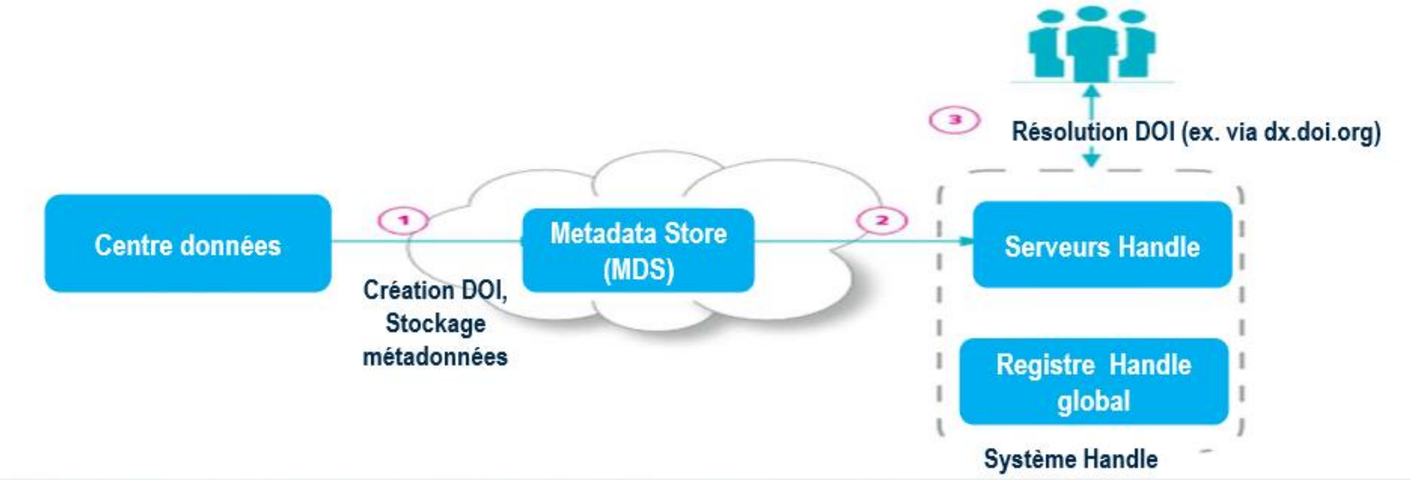
Infrastructure technique DataCite

Agence d'enregistrement Datacite

- Supporte l'infrastructure de résolution
- Maintient une base de métadonnées consultable et moissonnable
- Gère les identifiants sur le long terme

Utilisateurs (centres de donnée…)

- Assurance qualité
- Stockage et accessibilité du contenu
- Création des identifiants
- Création et mise à jour des métadonnées

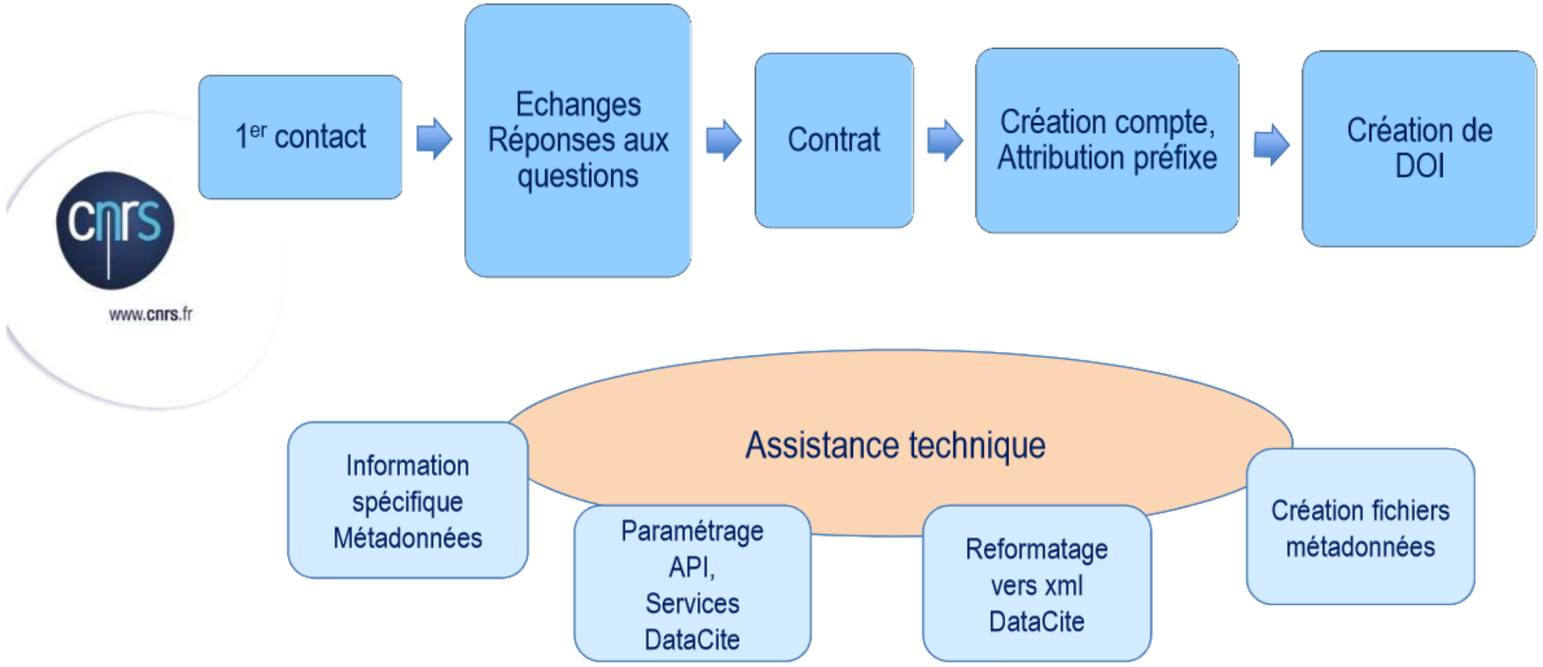


Services DataCite

Metadata Store (MDS)	Création des DOI Enregistrement des métadonnées associées	https://mds.datacite.org
Metadata schema	Schéma de métadonnées DataCite	https://schema.datacite.org
OAI Provider	Exposition des métadonnées de la base DataCite moissonnables selon le protocole OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)	https://oai.datacite.org
Metadata search	Moteur de recherche des jeux de données enregistrés dans la base DataCite	https://search.datacite.org

Attribution de DOI via l'INIST-CNRS

- L'INIST-CNRS est l'interlocuteur DataCite pour la France
- Contrat entre le centre de données et l'INIST
  - Forfait annuel de 180 € HT
  - Nombre illimité de DOI



Responsabilités des partenaires

INIST-CNRS

- Veiller au respect des règles de bonne pratique édictées par DataCite
  - Qualité des métadonnées
  - URL pérenne
  - Persistance des données
- Accompagner et conseiller à l'utilisation des services DataCite
- Fournir un préfixe de DOI unique
- Fournir un login pour accéder à la plateforme Metadata Store (MDS) de DataCite

Centre de données

- Garantir la qualité des données
- Garantir leur accessibilité sur le long terme
- Maintenir une page descriptive ou landing page accessible, contenant :
  - La citation
  - Les métadonnées descriptives
  - Les informations concernant l'accès à l'objet scientifique (URL, conditions d'obtention, restrictions, etc.)
  - Les informations pour lire l'objet scientifique (logiciels, contexte, autres informations nécessaires à l'interprétation….)
  - Eventuellement une information spécifiant l'indisponibilité des données

Attribution d'un DOI

10.6096/70765

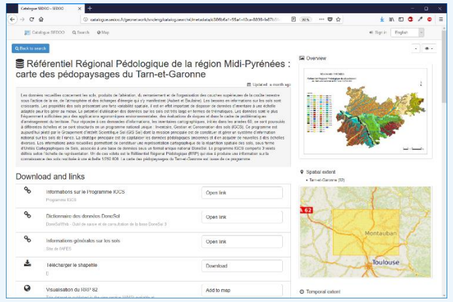
Préfixe — Suffixe

Nom du DOI

- Préfixe attribué par l'agence DataCite
- Suffixe choisi par le centre de données

Métadonnées

- Format XML
- 6 champs obligatoires



Landing page (URL)

- Descriptif des données
- Accès aux données

Métadonnées : 6 propriétés obligatoires

Table 1: DataCite Mandatory Properties

ID	***Property***	***Obligation***
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

- Ces informations permettent de générer la citation :

Creator (PublicationYear): Title. Publisher. Identifier

- Web service de formatage de la citation : https://citation.crosscite.org/

Métadonnées : 13 propriétés optionnelles

Table 2: DataCite Recommended and Optional Properties

ID	***Property***	***Obligation***
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O

Documentation complète :

https://schema.datacite.org/

Quelle granularité ?

- DataCite n'impose aucune restriction sur le niveau de granularité
- Il est possible d'attribuer des DOI à des entités de différents niveaux d'un même objet
- Les DOI sont principalement conçus pour la citation et la découverte de ressources.

  => Chaque entité devrait avoir un sens indépendant de l'ensemble plus large ou de la collection à laquelle elle peut appartenir
- Le Schéma de métadonnées DataCite comprend un champ pour préciser les relations entre les objets

Bonnes pratiques
(ateliers techniques inter-pôles)

Bonnes pratiques : Suffixe

- Le suffixe ne doit pas être porteur de sens.
- L'universalité du DOI doit être portée alors fois par le préfixe et le suffixe (et non par le suffixe seul)

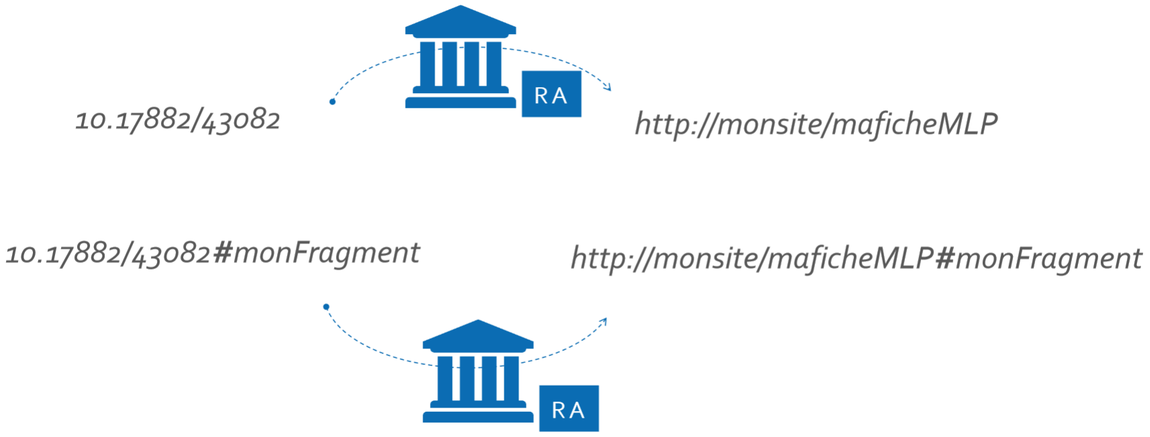
  ⇒ Permet d'avoir des identifiants plus concis

- Les variantes d'un jeu de données doivent être indiquées via des fragments et ne doivent pas donner lieu à la production de nouveaux DOI
- Les fragments doivent être concis et non significatifs.

  => Leur signification réelle doit être conservée et maintenue au sein de query stores

Fragments

- Norme W3C : https://www.w3.org/TR/media-frags/
- Supportés par les préconisations DOI : https://www.doi.org/doi_handbook/5_Applications.html#5.8



Bonnes pratiques : Landing Page

- La LP doit être une fiche de métadonnées structurée
    - Titre
    - Résumé
    - Contact(s)
    - Lien d'accès aux données
    - Etc.
- La LP doit proposer une citation afin de permettre son incorporation rapide dans un article
- La LP doit être capable d’interpréter un éventuel fragment

Bonnes pratiques : Producteur

- La production d'un DOI est une décision

  => Elle doit faire l'objet d'une décision concertée entre les différents acteurs de la donnée : équipes scientifiques, équipes techniques…

- Le producteur d'un DOI est le responsable de la conservation des données
- Le producteur doit mettre en place un mécanisme de vérification de l'état des LP
- Le fichier de données doit contenir un rappel des éléments de citation (au minimum son DOI)

- L'ensemble des métadonnées du DOI doivent provenir de manière automatisable de la fiche de métadonnées utilisée comme LP. Les métadonnées du DOI doivent être mise à jour si le contenu de la LP l'est.

Bonnes pratiques : résumé

1. **Un DOI est l'engagement d'une Landing Page pertinente pérenne**
2. Un DOI doit être:
    1. Réfléchi
    2. **Non significatif**
    3. **Complété par des fragments**
    4. Suivi par le responsable des données
3. La Landing Page doit être:
    1. **Riche** voire **intelligente**
    2. **Une fiche d'un catalogue**

Les DOI au SEDOO

Le SEDOO : Service de données de l'OMP

- Favoriser l'échange de données scientifiques via le développement d'applications informatiques
  - Gestion, traitement et archivage de données
  - Interfaces web d'accès aux données
  - Production de produits à valeur ajoutée

- Missions
  - Internationales : grands programmes multidisciplinaires (AMMA, MISTRALS), projets européens, etc.
  - Nationales : Pôles de données et de services, Services d'Observation (SO)
  - Locales : projets des laboratoires de l'OMP et du CNRM (AAP annuel)

- Plus de 30 applications en production
- Plus de 60 sites web institutionnels (projets, laboratoires, etc.)

Compétences

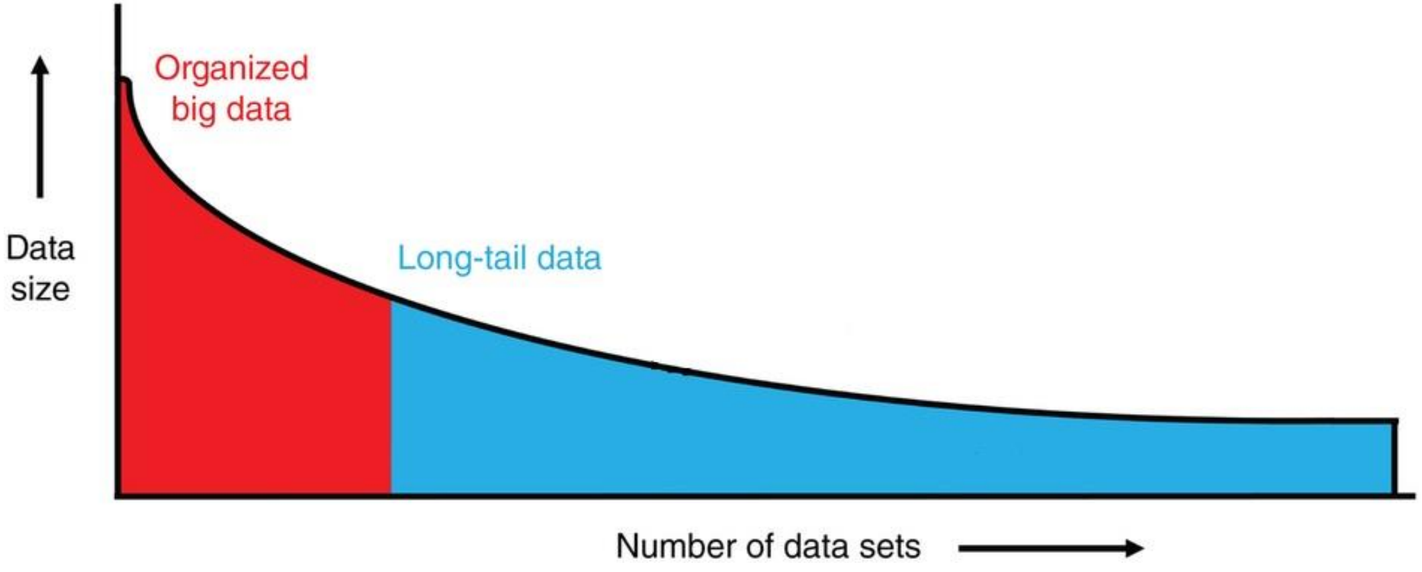
Compétences thématiques

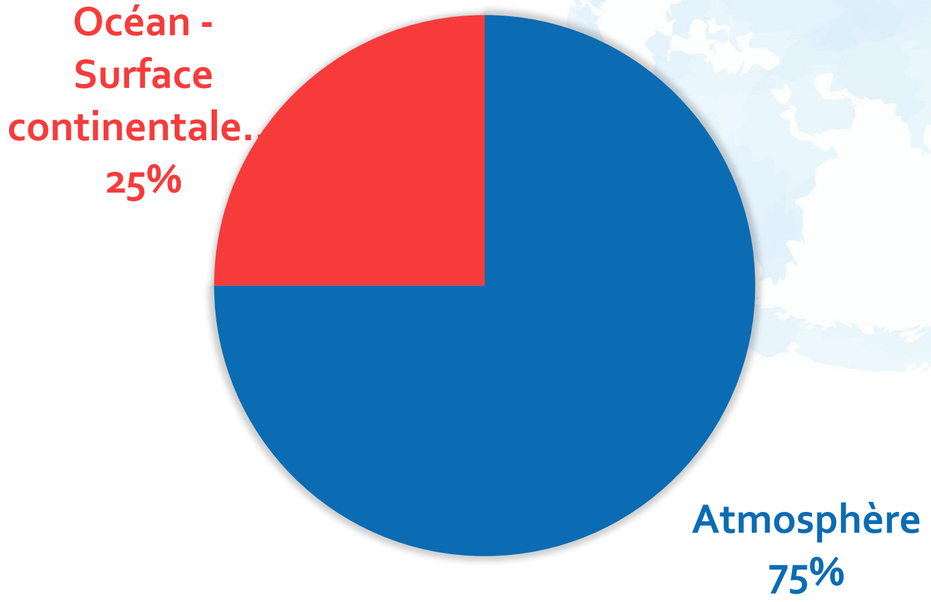
- Interaction avec les équipes de recherche
- Métadonnées (ISO, OGC, INSPIRE, Thesaurus)
- Attribution de DOI (Digital Object identifier)
- Homogénéisation de données (NASA Ames, NetCDF, etc.)
- Règles d'échange des données
- Promotion des outils (ateliers projets, conférences scientifiques)
- Support utilisateurs

Compétences techniques

- Conception, développement et exploitation de bases de données
- Traitement de données
- Développement web, ergonomie
- Web services, interopérabilité (OpenDAP, services OGC, etc.)
- Fonctionnalités avancées (visualisation, calcul, etc.)
- Suivi des utilisateurs, suivi des accès aux données
- Outils de soutien aux campagnes (sites temps réel, prévisions)

Données

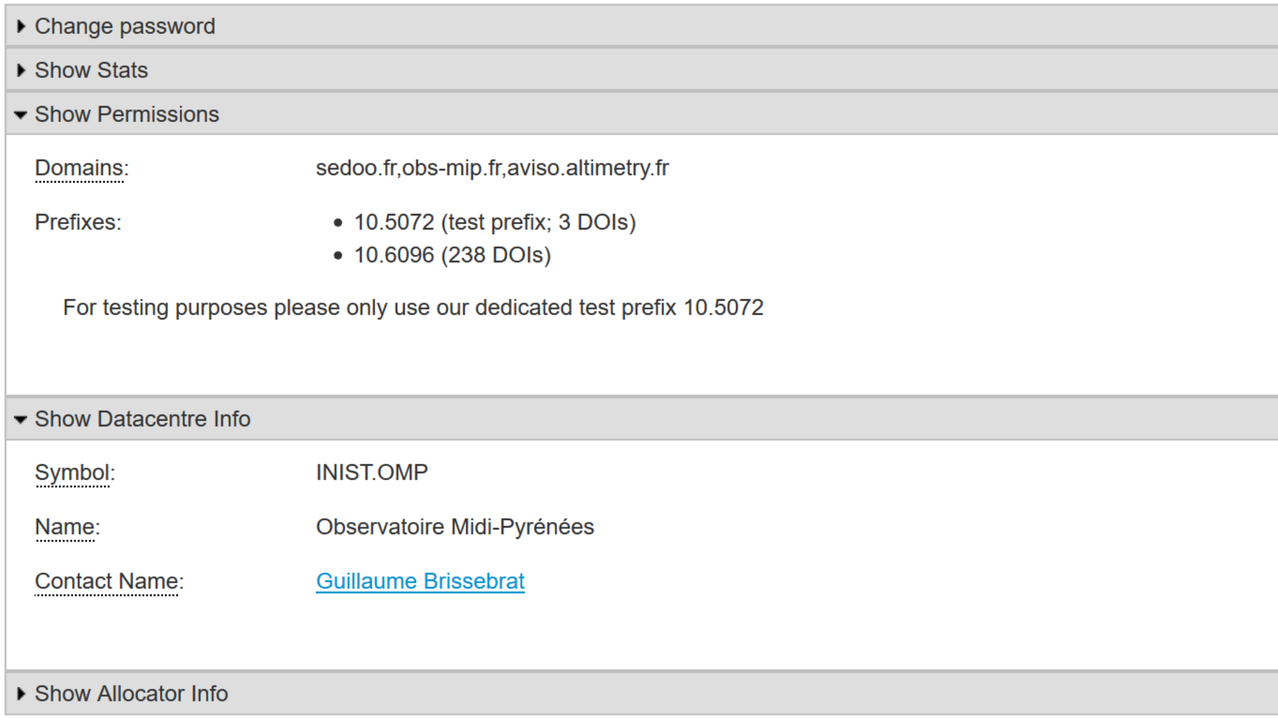




- Jeux de données hétérogènes
- Multidisciplinaires
- Volume de données faible (12 To au total)
- Stockage « sécurisé » grâce au service de sauvegarde et d'archivage du SIO (ex 2SAD)

Les DOI au SEDOO

- Convention avec l'INIST qui couvre l'OMP
- Attribution aux données gérées au SEDOO
  - Grands programmes (MISTRALS…)
  - SNO OMP (MSEC, BVET, SSS…)
  - Odatis
- Mais aussi à des données hébergées au CTOH
- Gestion du préfixe Aeris (pour les 4 CDS)
- Prochainement : un service de dépôt/publication de données pour l'OMP ?



Exemple : MISTRALS

General information

Dataset name	ATR AVIRAD nephelometer - LISA - TRAQA
DOI	**10.6096/MISTRALS-ChArMEx.1012** **Citation:** Formenti Paola and Claudia, D. B. (2013) 'AVIRAD nephelometer - TRAQA'. SEDOO OMP. doi: 10.6096/mistrals-charmex.1012. BibTeX
Created on	2013-09-27
Project(s)	ChArMEx > TRAQA
Period	SOP-0 summer 2012
Contacts	**Formenti Paola - LISA (PI or Lead scientist)**
Data access	**Dataset as provided by the Principal Investigator**
History	ISSUE **2013-10-03**

Data description

Abstract	Aerosol scattering and backscattering coefficient at 3-wavelengths
Observing strategy	Instrument in the ATR-42, part of the AVIRAD sampling system

Instrument information

Instrument type	NEPHELOMETERS
Manufacturer	TSI Inc. - **http://www.tsi.com/**
Model	3563
Observation frequency	5 sec

Geographic information

Dataset	AEROSOL CCNC CNRM/GAME ATR42/SAFIRE SOP1
DOI	10.6096/MISTRALS-HyMeX.1005

Metadata (xml)	`<?xml version="1.0" encoding="UTF-8"?>`
	`<!--Sample XML file generated by XMLSpy v2005 rel. 3 U (http://www.altova.com)-->`
	`<resource xmlns="http://datacite.org/schema/kernel-3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://datacite.org/schema/kernel-3 http://schema.datacite.org/meta/kernel-3/metadata.xsd">`
	`  <identifier identifierType="DOI">10.6096/MISTRALS-HyMeX.1005</identifier>`
	`  <relatedIdentifiers>`
	`  <relatedIdentifier relatedIdentifierType="URL" relationType="References">http://mistrals.sedoo.fr</relatedIdentifier>`
	`  <relatedIdentifier relationType="References" relatedIdentifierType="URL">http://www.hymex.org/</relatedIdentifier>`
	`  </relatedIdentifiers>`
	`  <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>`
	`  <publisher>SEDOO OMP</publisher>`
	`  <creators>`
	`  <creator>`
	`  <creatorName>Roberts Greg</creatorName>`
	`  </creator>`
	`  <creator>`

Error	

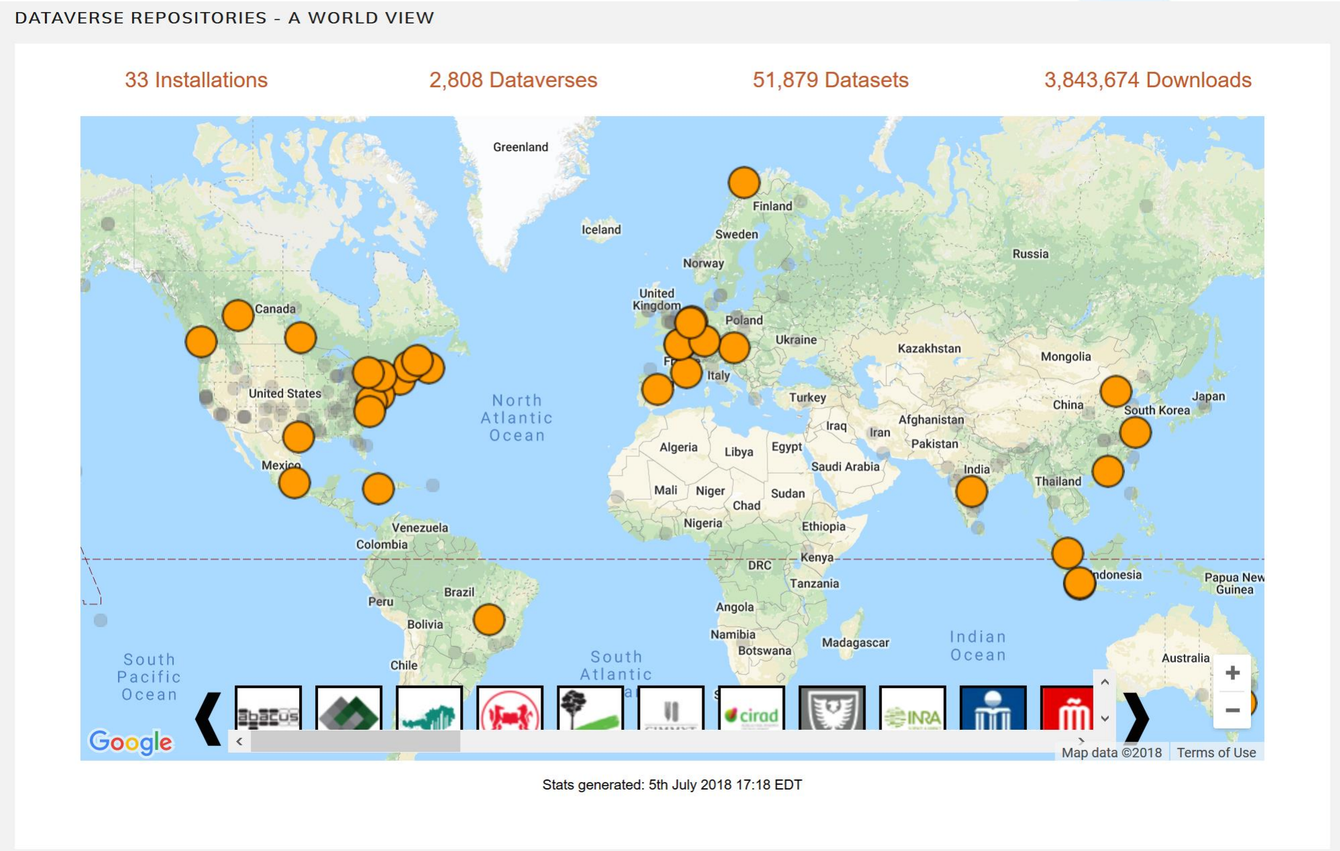
Create DOI

- Métadonnées générées automatiquement à partir de la base de données MISTRALS
- Création du DOI via l'API DataCite

https://support.datacite.org/docs/mds-api-guide

Dataverse

- Un logiciel open source et gratuit dédié à la gestion et au partage des données scientifiques
- Créé en 2006 par l'Institute for Quantitative Social Sciences en collaboration avec Harvard Library et Harvard University Information Technology
- Accueille tous formats de fichiers
- Permet de créer un réseau de portails



Fonctionnalités

Features

Data Citation
automatically generated

Multiple Publishing Workflows
dataset in draft, in review, and then published

Terms of Use + Guestbook
CC0 waiver default, custom terms of use, and download metrics

Account + Data Notifications
access request, roles granted, and when data is published to name a few

Faceted Search
metadata fields based facets

Pull header metadata from Astronomy (FITS) files

APIs for interoperability
search API, data deposit API

Shibboleth
single sign on using your institution's credentials

Three Levels of Metadata
description/citation, domain-specific or custom fields, file metadata

Access Control Support
pre-defined and custom roles

Restricted Files + Ability to request access to restricted files
allow anyone, certain people, or no one to be able to download files

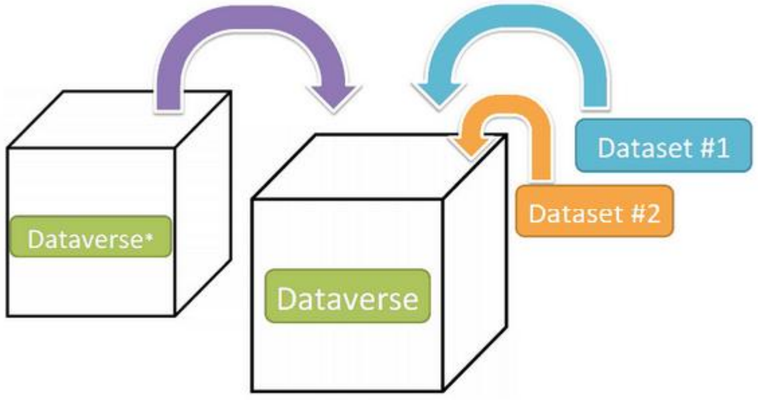
Customization of dataverses
branding, metadata based facets, sub-dataverses, featured dataverses

Re-format, Summary Statistics, and Analysis for Tabular Files
integration with TwoRavens

Mapping of Geospatial files
integration with WorldMap

Fonctionnement

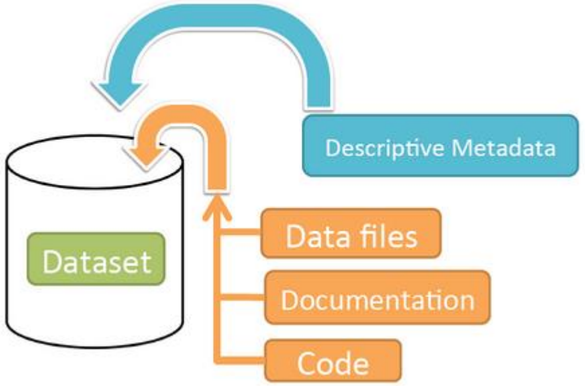
Schematic Diagram of a **Dataverse** in Dataverse 4.0



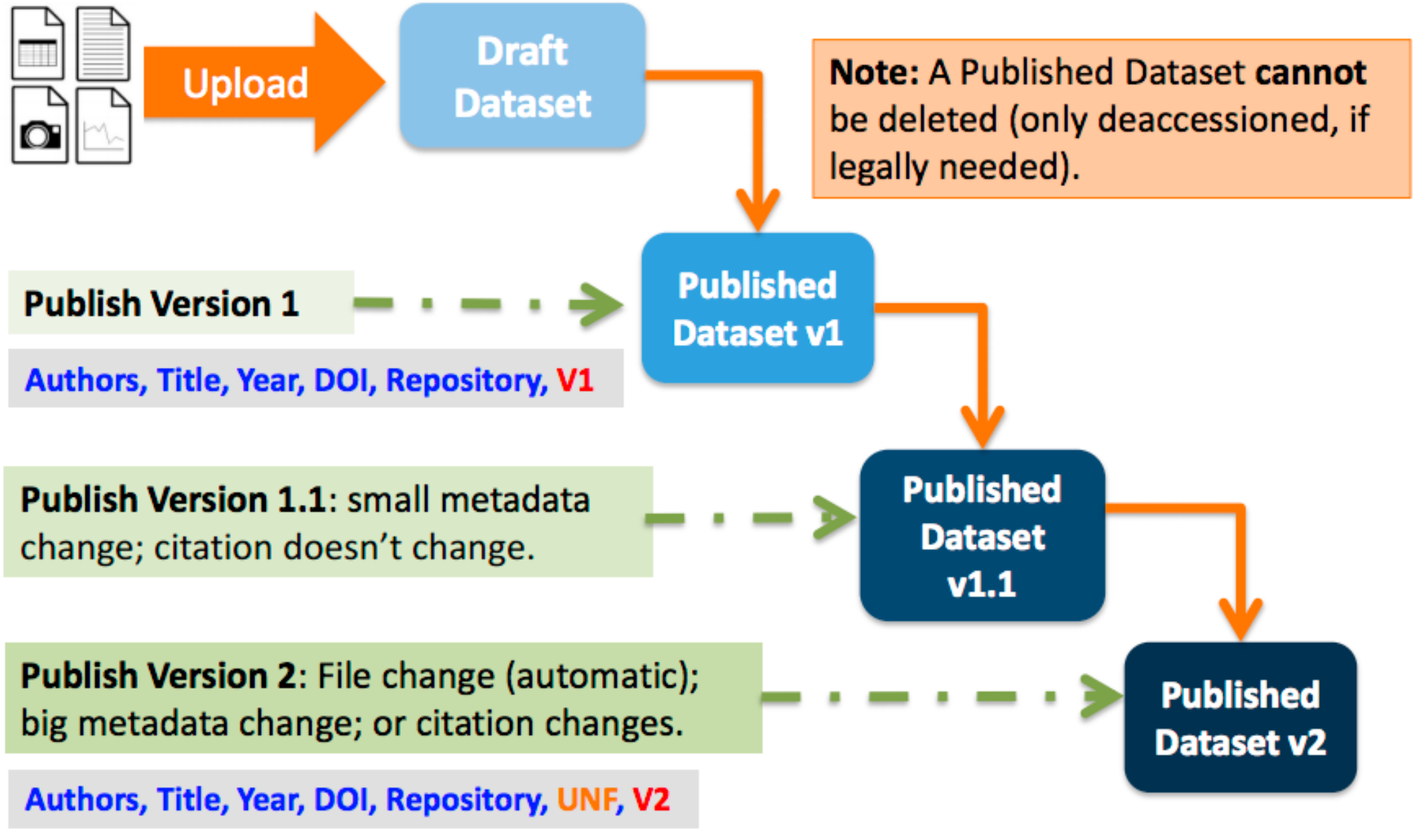
Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

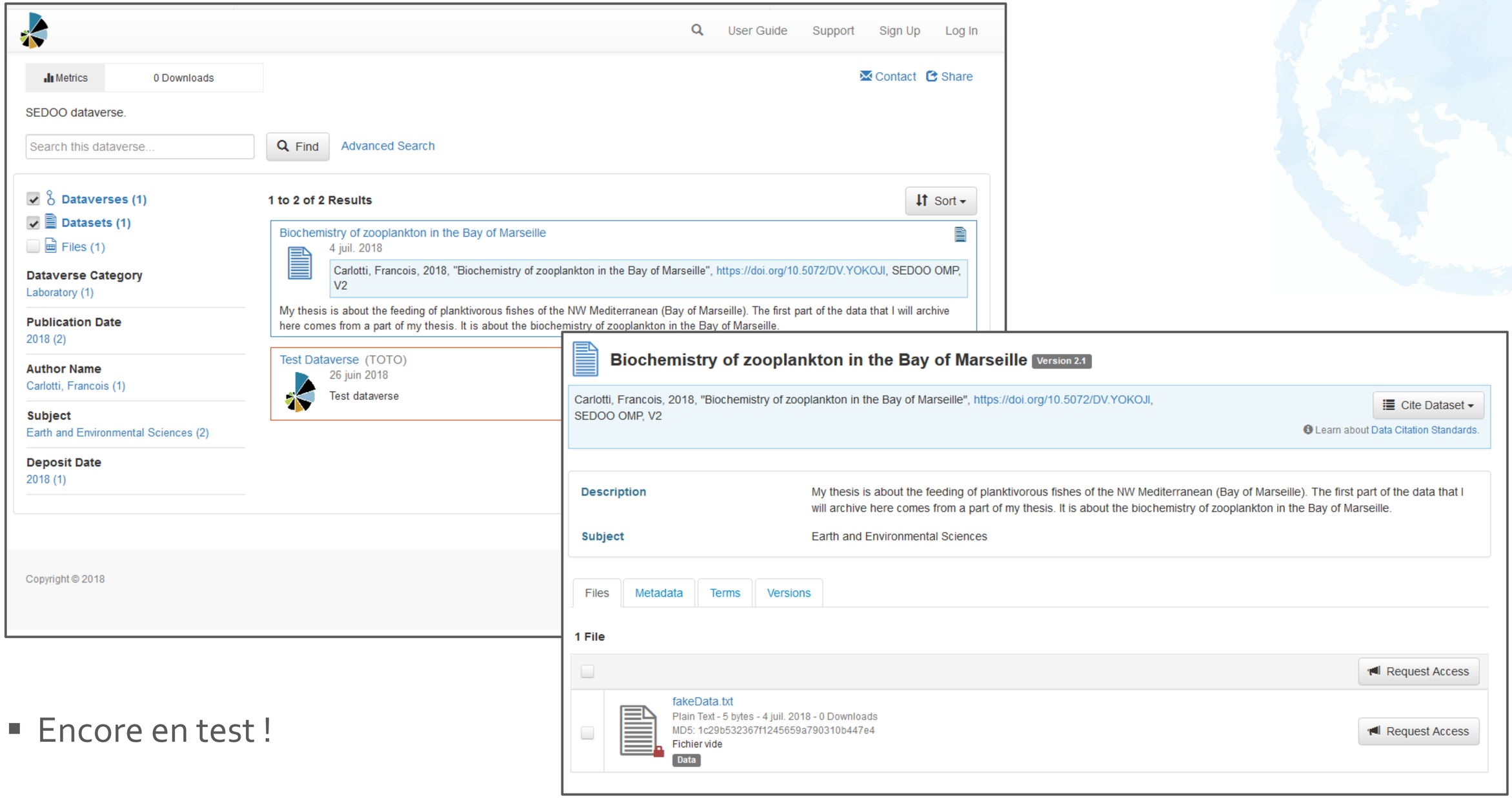
Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.



Dataverse SEDOO OMP



- Encore en test !

Merci