Les entrepôts de données de recherche

Sylvie Cocaud, Pascal Aventurier

Les entrepôts de données de recherche

Action Nationale de Formation RENATIS – MEDICI 2017 (3-6/07/2017) :
Participer à l'organisation du management des données de la recherche, gestion de contenu et documentation des données

Programme

- Qu’est ce qu’un entrepôt ?
- Quelques exemples
- Choisir un entrepôt
- Conclusion

Qu’est ce qu’un entrepôt ?

Sylvie Cocaud

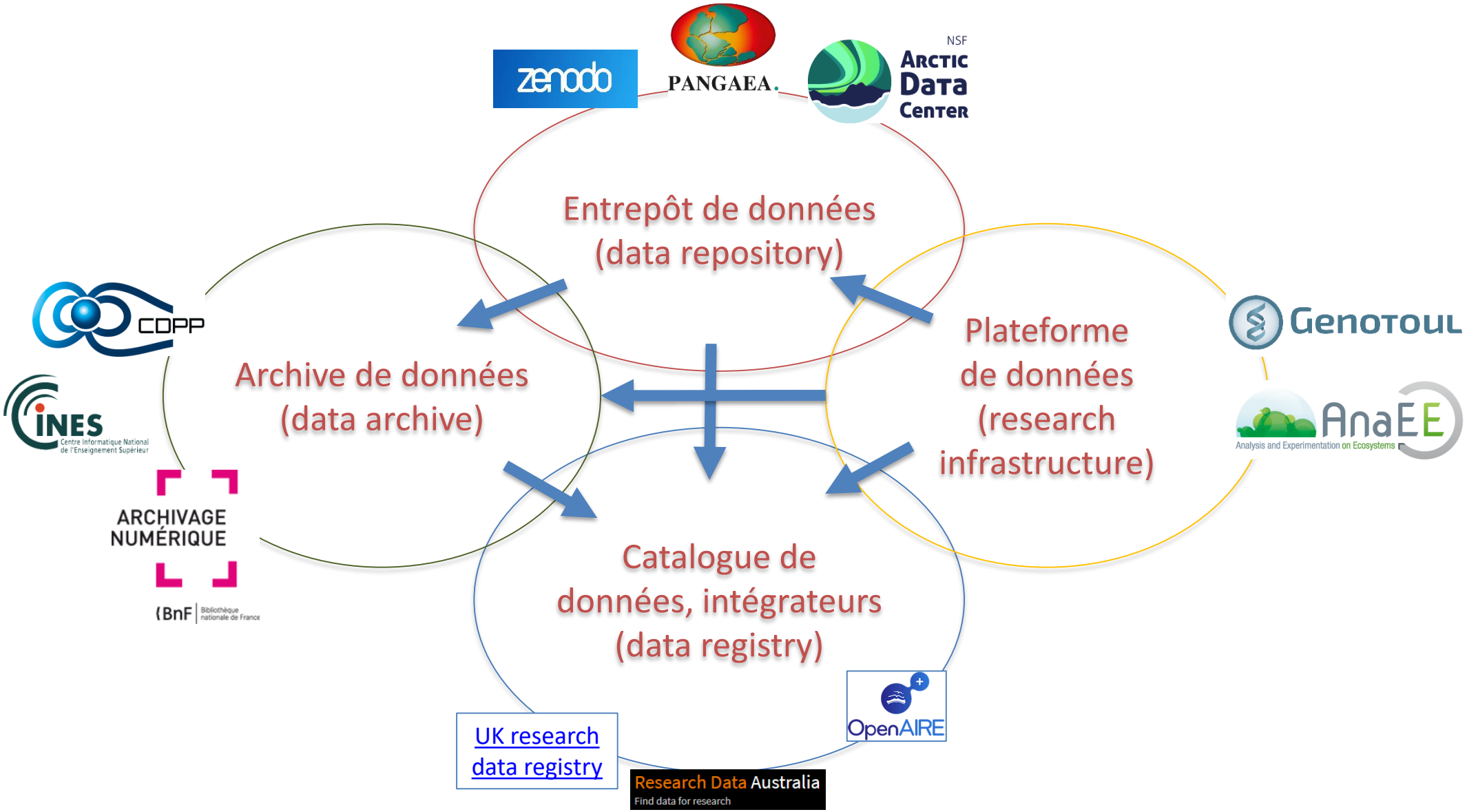
Les entrepôts de données de recherche

ist@inra

INRA
SCIENCE & IMPACT

Définition : Entrepôt de données de recherche

Service en ligne permettant la collecte, la description, la conservation, la recherche et la diffusion des jeux de données.


zenodo
PANGAEA.
NSF
Arctic Data Center
Entrepôt de données
(data repository)
CDPP
INES
Centre Informatique National de l'Enseignement Supérieur
Archive de données
(data archive)
ARCHIVAGE NUMÉRIQUE
BnF Bibliothèque nationale de France
Plateforme
de données
(research
infrastructure)
Genotoul
AnaEE
Analysis and Experimentation on Ecosystems
Catalogue de
données, intégrateurs
(data registry)
UK research data registry
OpenAIRE
Research Data Australia
Find data for research

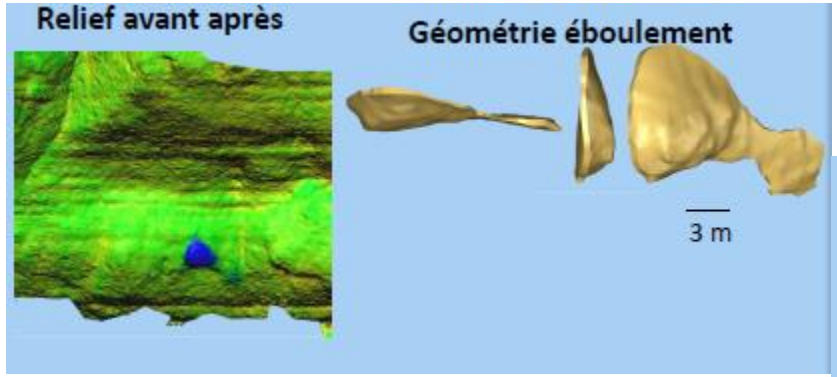
Définition : données de la recherche (research data)

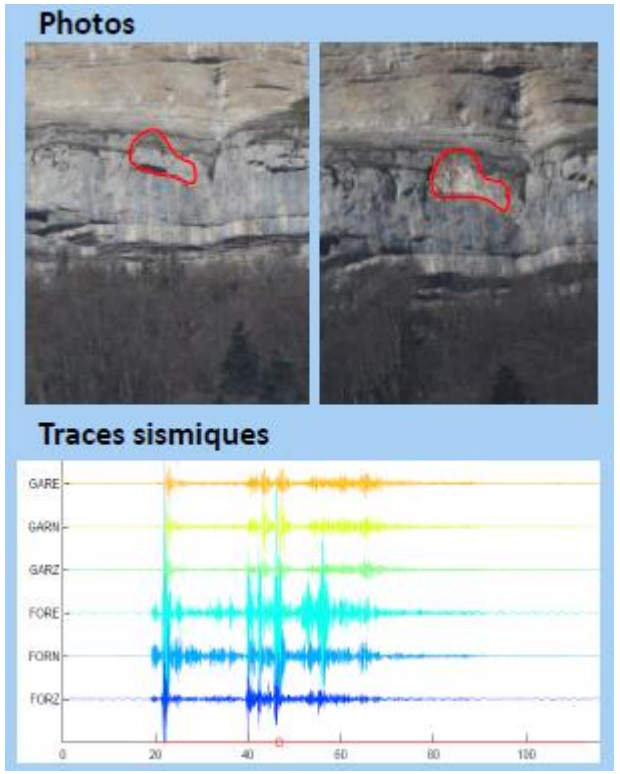
- « Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires à la validation des résultats de recherche » (OCDE, 2007)
- « Les données de la recherche sont l'ensemble des informations et matériaux produits et reçus par des équipes de recherche et des chercheurs. Elles sont collectées et documentées à des fins de recherche scientifique. À ce titre, elles constituent une partie des archives de la recherche » (Pomart, 2014)

Définition : Données brutes - données élaborées (raw data – processed data)

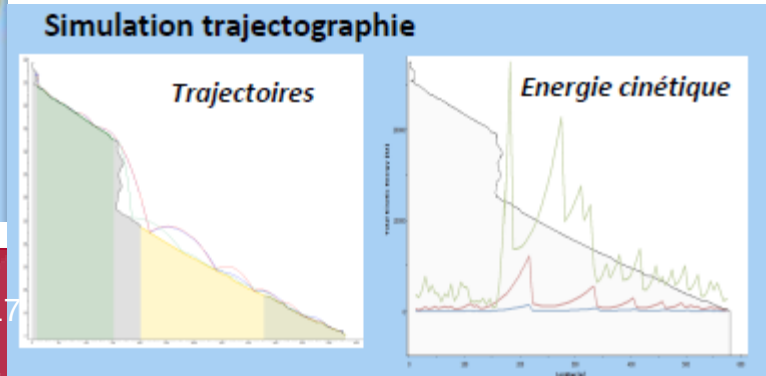
- « Données brutes : recueillies sur un sujet à partir d'observations ou de mesures, et qui n'ont pas encore été traitées, validées ou triées. Sensées être neutres et objectives et ne dépendant pas de leur contexte de création, d'une analyse ou de leur producteur » (Thessen & Patterson, 2011)

- Données élaborées ou dérivées : ayant subi des opérations de nettoyage, de sélection, de traitement, d'analyse afin de produire des résultats.





(Amitrano, Lejeune, & Hobléa, 2017)



Définition : Métadonnée (metadata)

- **Donnée qui renseigne sur la nature d'autres données et qui permet ainsi leur utilisation pertinente.**
  - Dans la perspective des entrepôts de données, les métadonnées sont un **élément primordial** et sont destinées à diverses catégories d'utilisateurs. Elles permettent notamment de connaître **l'origine et la nature des données** stockées dans l'entrepôt, de comprendre **comment elles sont structurées,** de savoir **comment y avoir accès** et **comment les interpréter,** de connaître les différents **modèles de données** en présence et les **règles de gestion** de ces données (Office québécois de la langue française, 2002).
    - Certaines métadonnées sont générées dès la création des données, automatiquement ou manuellement. Fournir les métadonnées le plus rapidement possible après la création des données.

Définition : Jeu de données (dataset)

- **Agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine "unité", rassemblées pour former un ensemble cohérent (Gaillard R., 2014).**
- **Certains entrepôts utilisent le terme « data package » plutôt que « dataset » pour désigner un ensemble constitué des fichiers de données et des métadonnées associées**
  - Le nombre de fichiers peut être très important dans un jeu de données

citable comme un objet unique

Fonctionnalités des entrepôts : dépôt et conservation des données

Téléchargement / import de fichiers :

- données + documents complémentaires
- Limites / taille des fichiers, Formats acceptés, données publiées…
- Via IU ou API

Organisation des données en collections

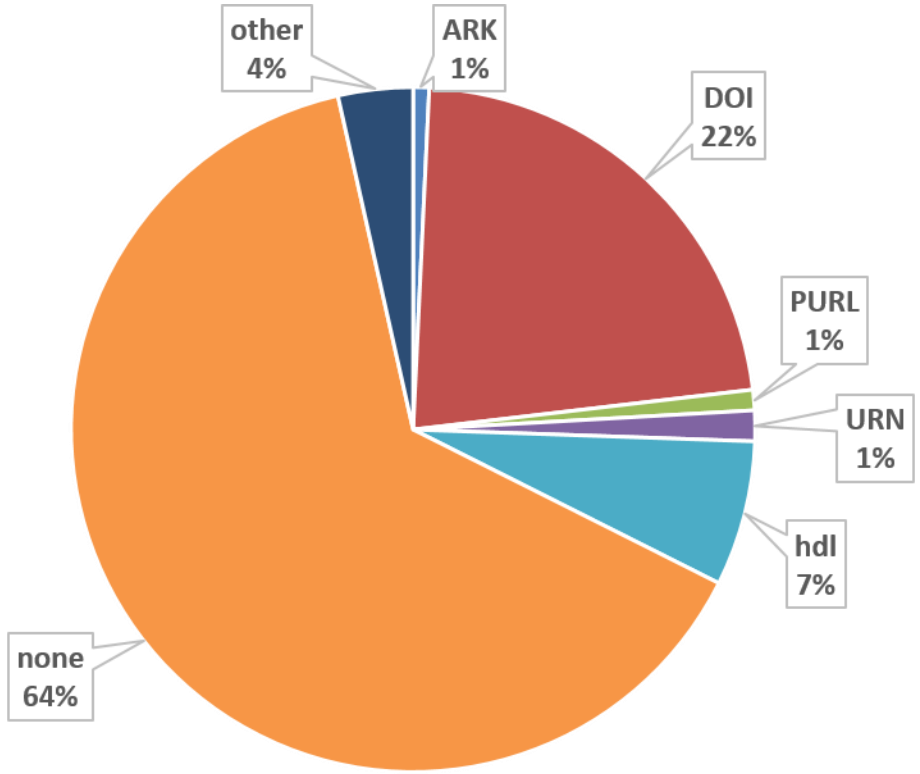
Conservation (stockage sécurisé + archivage à long terme)

Intégrité	• sauvegarde • empreinte numérique
Lisibilité	• veille sur l'évolution des technologies, • utilisation de formats durables (ou conversion), standards
Intelligibilité	• métadonnées riches, standards • identification pérenne • contrôle des versions de jeux de données.

ist@inra

Fonctionnalités des entrepôts : identification pérenne des données

- **Attribution d'un identifiant au moment du dépôt**
  - Au jeu dans son ensemble,
  - +/- à chaque version du jeu de données
- **Intégration possible d'identifiant préexistant chez certains entrepôts (ex : Zenodo)**



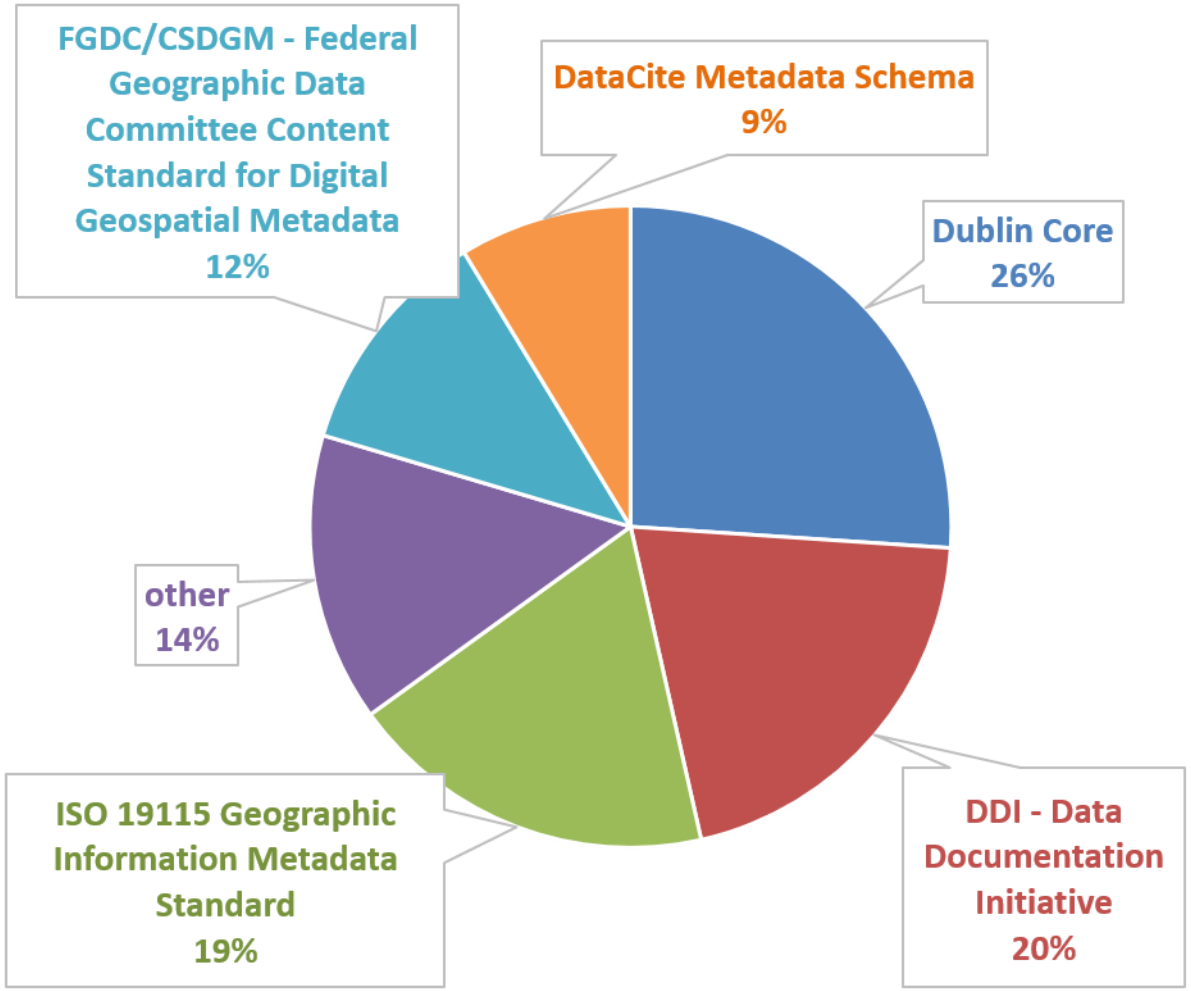
Données extraites de re3data le 19/05/2017

Fonctionnalités des entrepôts : description des données

Métadonnées

- Schémas de métadonnées
  - génériques
  - spécifiques à un domaine
- Qualité des métadonnées
  - vocabulaires contrôlés (import ou accès à des vocabulaires externes)
- Saisie import via IU vs API

Documentation complémentaire



Données extraites de re3data le 19/05/2017

Exemple de métadonnées

Datacite v 4.0

- **Identifier**
- **Creator**
- **Title**
- **Publisher**
- **PublicationYear**
- **ResourceType**
- Subject
- Contributor
- Date
- Language
- AlternateIdentifier
- RelatedIdentifier
- Size
- Format
- Version
- Rights
- Description
- GeoLocation
- FundingReference

CIF (cristallographie) : extrait, ex. de valeurs

a	10.4686 ± 0.0016 Å
b	11.9543 ± 0.0016 Å
c	5.9271 ± 0.0016 Å
α	102.365 ± 0.016°
β	92.12 ± 0.016°
γ	65.302 ± 0.015°
Cell volume	657.1 ± 0.2 Å^3
Cell temperature	293 ± 2 K
Ambient diffraction temperature	293 ± 2 K
Number of distinct elements	4
Hermann-Mauguin symmetry space group	P -1
Hall symmetry space group	-P 1
Residual factor for all reflections	0.0727
Residual factor for observed reflections	0.0419
Weighted residual factors for the observed reflections	0.1149
Weighted residual factors for all reflections included in the refinement	0.1333
Goodness-of-fit parameter for observed reflections	1.05
Diffraction radiation wavelength	0.71069 Å
Diffraction radiation type	MoKα
Has coordinates	Yes
Has disorder	Yes
Has F_{obs}	No

Fonctionnalités des entrepôts : contrôle des droits d'accès aux données, conditions d'utilisation et licences

Droit d'accès aux données :

- **téléchargement libre (open)**
- embargo
- demande d'accès (restricted), guestbook
- pas de téléchargement possible (closed)
  - Génération d'URL privé parfois proposé

Attribution d'une licence à chaque jeu de données

- Saisie libre vs liste fermée
  - attention aux entrepôts qui imposent une licence unique

ist@inra

Fonctionnalités des entrepôts : recherche, affichage, export des (méta)données

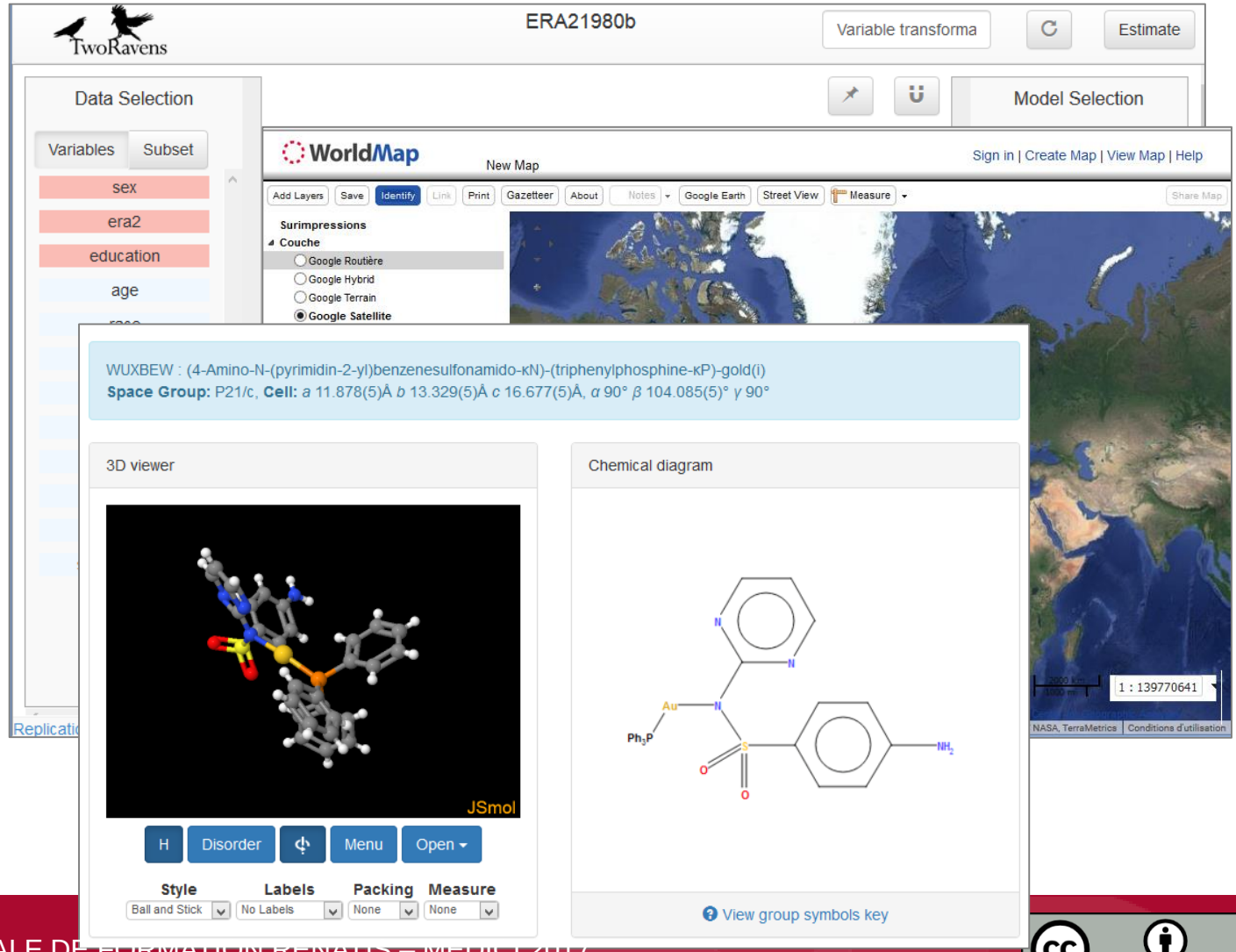
- **Recherche**
  - simple et/ou avancée : dans les métadonnées, parfois dans les données,
  - Graphique (zone géographique, structure molécule…)
  - Navigation et affinage par facettes
- **Affichage**
  - des métadonnées
    1. Liens avec d'autres datasets, avec des publications
    2. Génération de la citation
  - des données (prévisualisation, selon format)
- **Exports**
  - métadonnées (≠ formats),
  - téléchargement des données
- **Génération de template de data papers à partir des métadonnées**

ist@inra

Fonctionnalités des entrepôts : exploration et visualisation des données

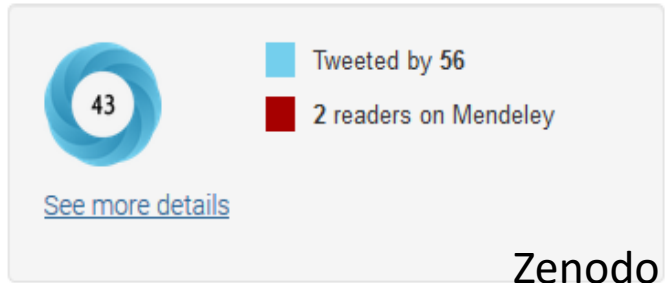
Outils d'analyse ou de visualisation

- TwoRavens,
- WorldMap,
- Mercury
- …



Fonctionnalités des entrepôts : découverte, visibilité des données

- **Exposition des métadonnées via OAI-PMH ou dans un triple store RDF (Linked data)**
- **API pour la recherche et l'accès**
- **Liaison avec d'autres ressources :**
  - Ex : IsCitedBy, Cites, … qui sont les valeurs possibles de la propriété relationType de la métadonnée relatedIdentifier
- **Statistiques d'usage de ses données**
  - nombre d'affichages et de téléchargements, datasets les + téléchargés, altmetrics, …



Recherche à partir des métadonnées DataCite

DataCite propose une interface et une API de recherche : exploration des métadonnées possibles quand un DOI est attribué à un jeu de données
Exemple : chercher les ressources génétiques identifiées par l'INRA

https://search.datacite.org/api?&q=prefix:10.15454 AND resourceType:« Genetic Resource »
&fl=doi,creator,title,publicationYear,rights,resourceTypeGeneral,resourceType
&wt=csv
&rows=100

Tester

Texte libre vs vocabulaire

doi,creator,title,publicationYear,rights,resourceTypeGeneral,resourceType
10.15454/1.49380688032889l6E12,GnpIS,INRA:Pinus pinaster:M421T-118,2017,CC BY,PhysicalObject,Genetic Resource
10.15454/1.49380688032424l3E12,GnpIS,INRA:Pinus pinaster:M421T-119,2017,CC BY,PhysicalObject,Genetic Resource
10.15454/1.49380688032413l74E12,GnpIS,INRA:Pinus pinaster:M421T-12,2017,CC BY,PhysicalObject,Genetic Resource
10.15454/1.49380688032494l78E12,GnpIS,INRA:Pinus pinaster:M421T-120,2017,CC BY,PhysicalObject,Genetic Resource

Fonctionnalités des entrepôts : gestion des utilisateurs et des droits

Authentification

- Intégration de systèmes externe (fédération d'identifié, comptes ORCID, GitHub…)

Gestion des utilisateurs et des droits

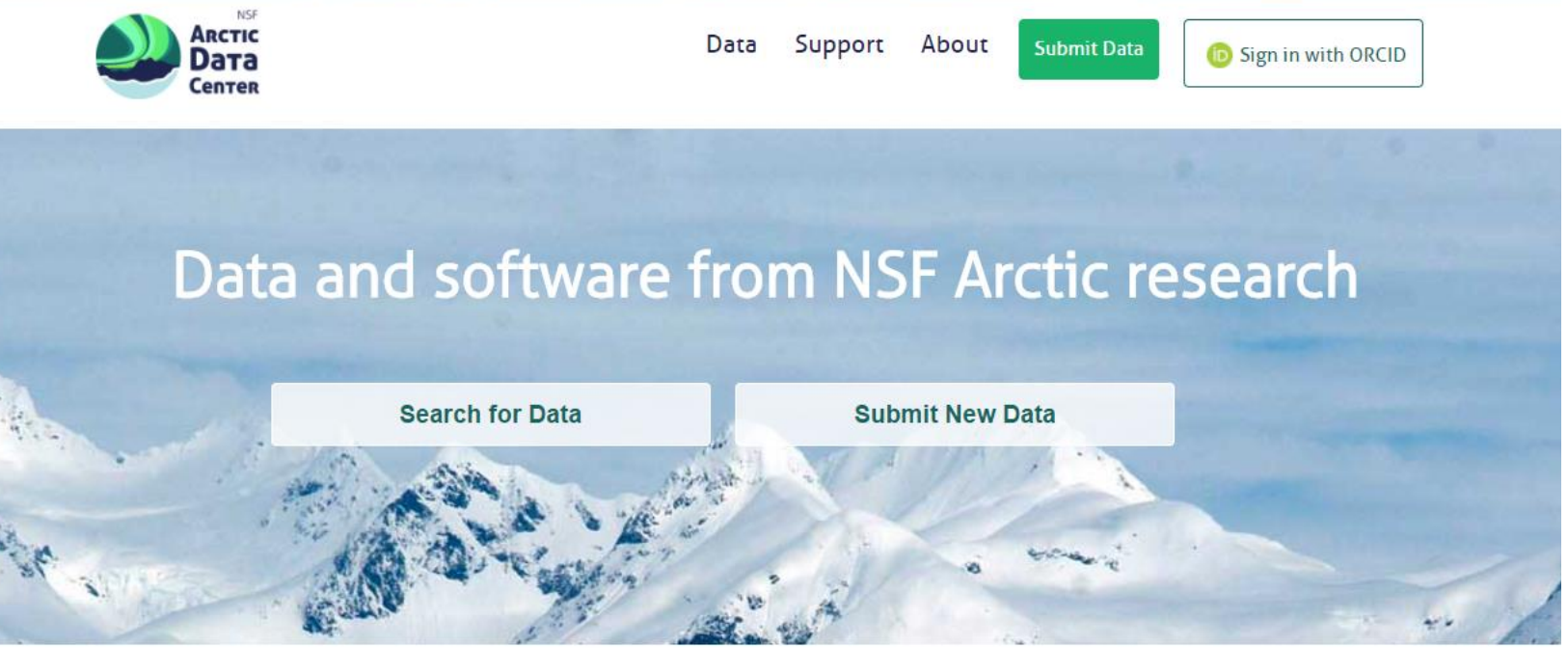
- Gestion de rôles (administrateur de collection, curateur…)
- Attribution de droits d'accès à des utilisateurs, à des groupes d'utilisateurs (collection, communautés, …)
- Workflow de publication des données

Quelques exemples

Pascal Aventurier

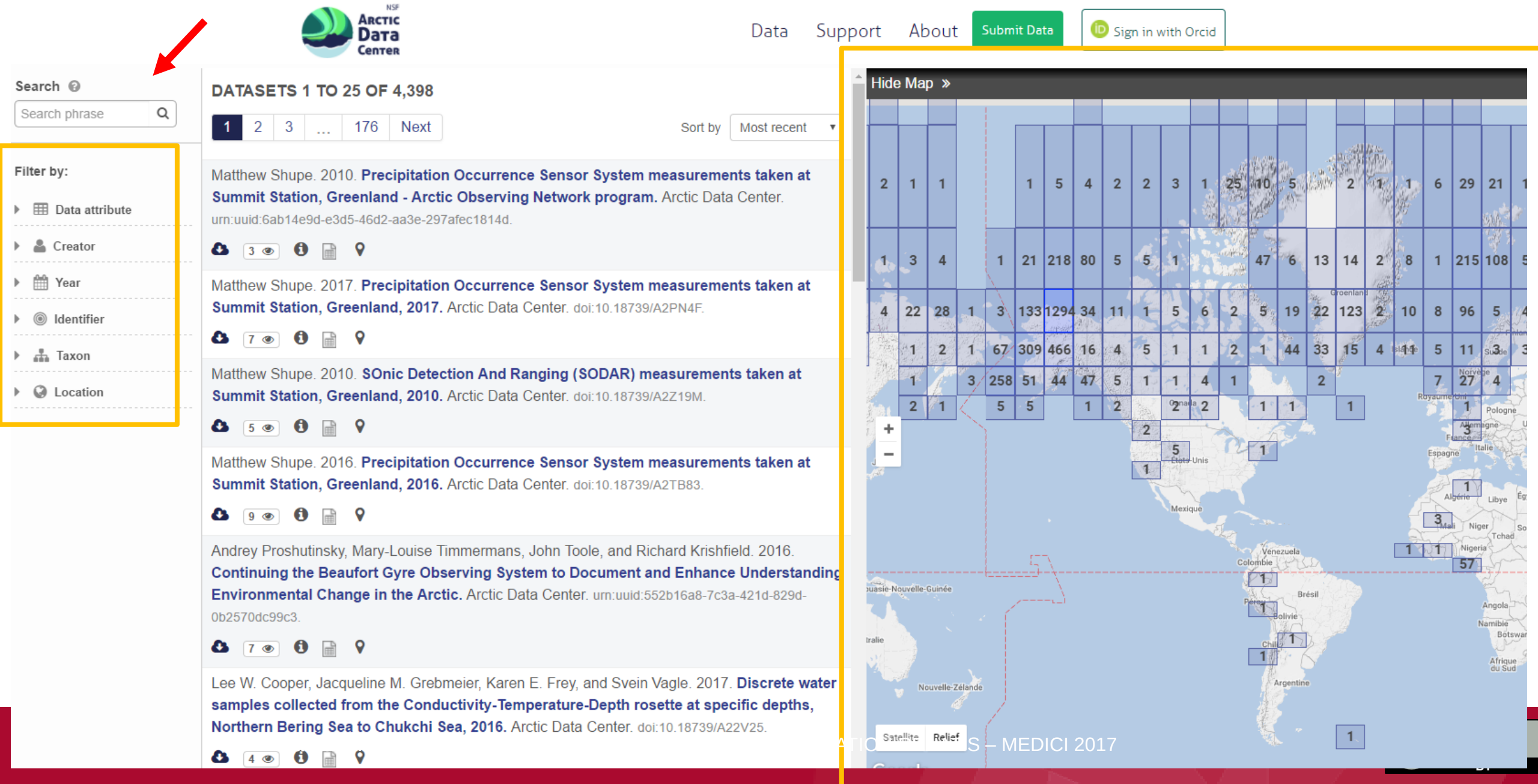
Les entrepôts de données de recherche

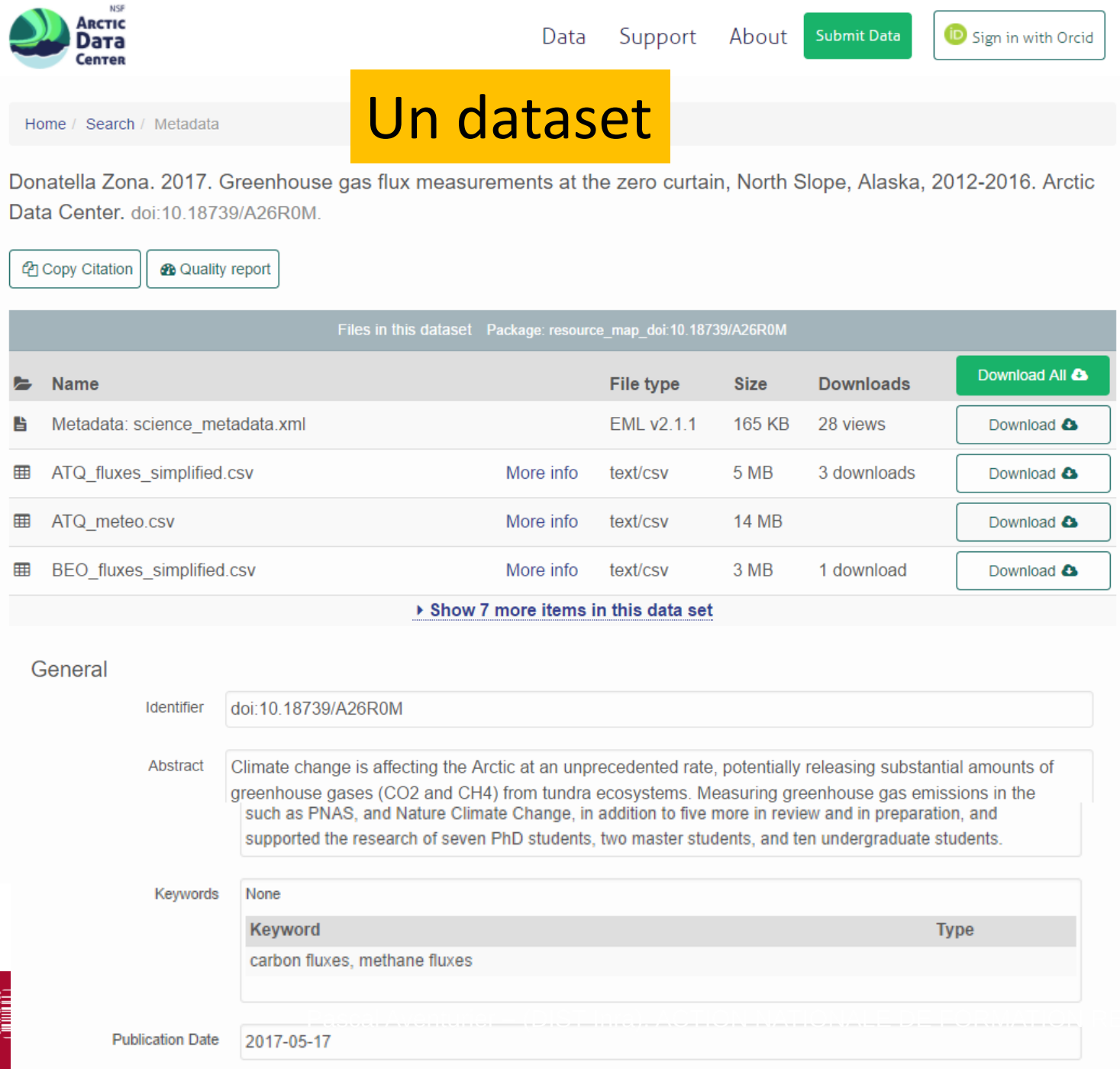
https://arcticdata.io



- Créé en en 2007, entrepôt de la NSF depuis 2016
- Métadonnées : EML Ecological Metadata Language
- Description des différentes variables
- Entrepôt « Nœud » de DataOne pour la réplication des données
- Dépôt possible par une interface web ou par programme R/Matlab

Recherche


NSF Arctic Data Center
Data
Support
About
Submit Data
Sign in with Orcid
Search
Search phrase
Filter by:
Data attribute
Creator
Year
Identifier
Taxon
Location
DATASETS 1 TO 25 OF 4,398
1 2 3 ... 176 Next
Sort by Most recent
Matthew Shupe. 2010. Precipitation Occurrence Sensor System measurements taken at Summit Station, Greenland - Arctic Observing Network program. Arctic Data Center. urn:uuid:6ab14e9d-e3d5-46d2-aa3e-297afec1814d.
Matthew Shupe. 2017. Precipitation Occurrence Sensor System measurements taken at Summit Station, Greenland, 2017. Arctic Data Center. doi:10.18739/A2PN4F.
Matthew Shupe. 2010. SOnic Detection And Ranging (SODAR) measurements taken at Summit Station, Greenland, 2010. Arctic Data Center. doi:10.18739/A2Z19M.
Matthew Shupe. 2016. Precipitation Occurrence Sensor System measurements taken at Summit Station, Greenland, 2016. Arctic Data Center. doi:10.18739/A2TB83.
Andrey Proshutinsky, Mary-Louise Timmermans, John Toole, and Richard Krishfield. 2016. Continuing the Beaufort Gyre Observing System to Document and Enhance Understanding Environmental Change in the Arctic. Arctic Data Center. urn:uuid:552b16a8-7c3a-421d-829d-0b2570dc99c3.
Lee W. Cooper, Jacqueline M. Grebmeier, Karen E. Frey, and Svein Vagle. 2017. Discrete water samples collected from the Conductivity-Temperature-Depth rosette at specific depths, Northern Bering Sea to Chukchi Sea, 2016. Arctic Data Center. doi:10.18739/A22V25.
Hide Map »
Satellite
Relief



Fichiers de données

Métadonnées 1

Métadonnées 2

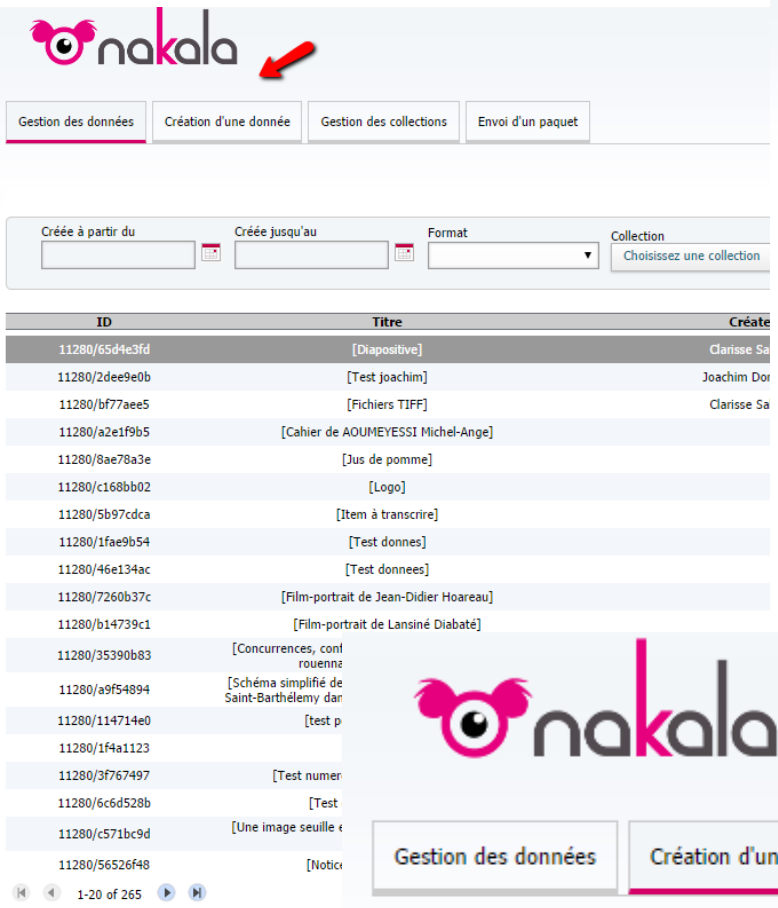
Métadonnées très détaillées
Avec les attributs du jeu de données

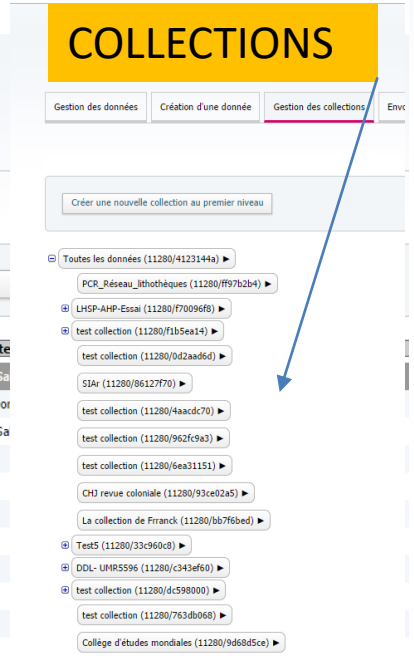
https://www.nakala.fr/nakala/

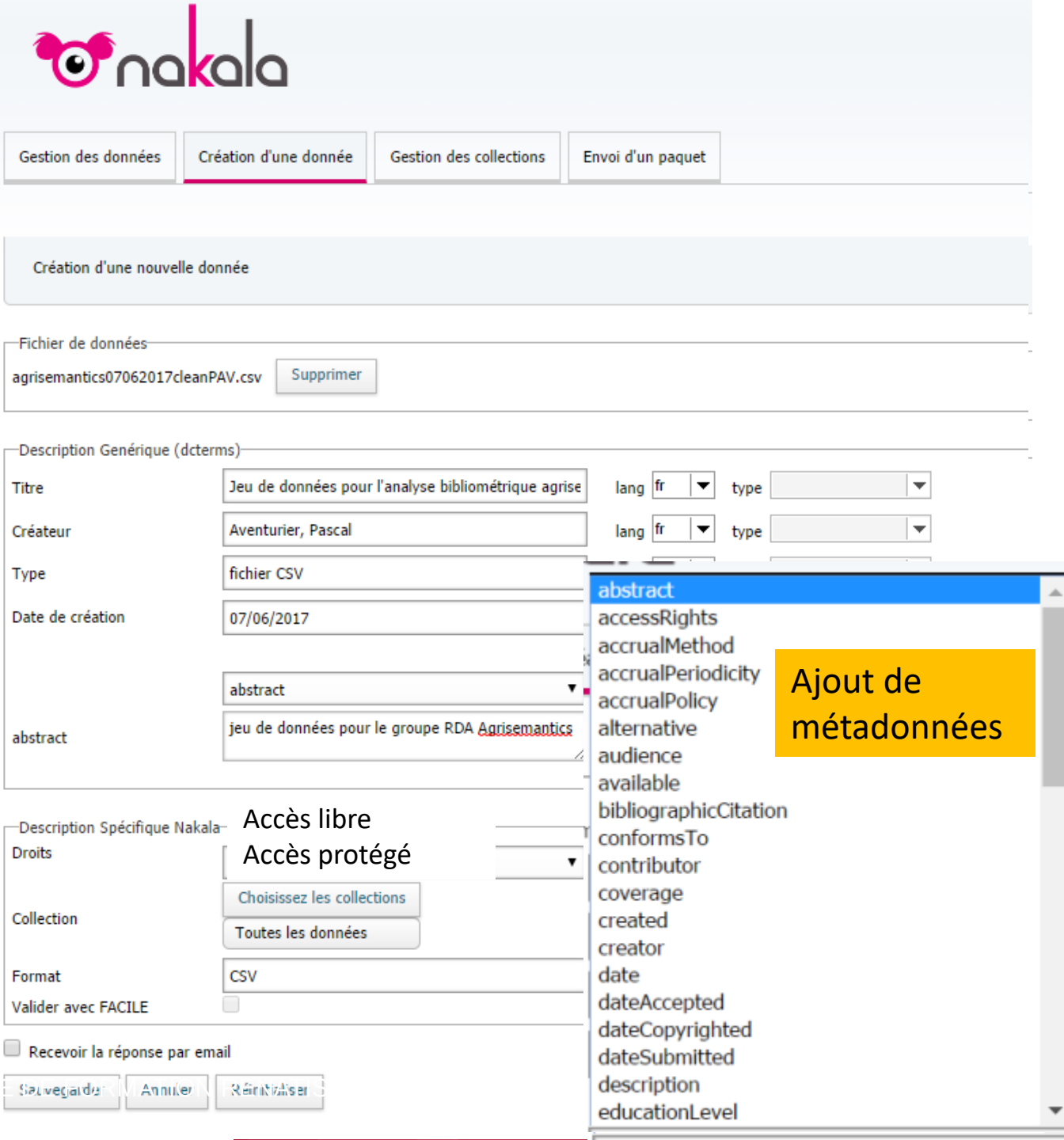
- Création Septembre 2014, par TGIR Huma-num
- Contenu. Fichiers texte, sons, images, vidéo, discipline SHS.
- Ouvert en dépôts aux unités de recherche qui en font la demande
- Outil pour déposer, documenter et Diffuser les données de la recherche
- Fonctionnalités intéressantes
  - pas d'interface de recherche dédiée
  - Possibilité de connecter les métadonnées à d'autres entrepôts existants, et de les rendre moissonnables par des services spécialisés comme *ISIDORE*.
  - Diffusion des données vidéos, images dans les carnets hypothèses*
  - Exposition des données ou des métadonnées
  - Sparql Endpoint pour interroger les métadonnées
  - Création de collection
  - Identifiant handle peut être adressé dans le web
  - accompagnement aux chercheurs via les consortiums HumaNum

*https://humanum.hypotheses.org/774

Dépôt dans Nakala







- Création
  - Projet EU-H2020 créé en 2011 , 2eme phase EUDAT2020 jusqu'en 2018

- Contenu

Plus de 500 jeux de données, moissonne pus de 450 000 métadonnées.

- Fonctionnalités intéressantes
  - Des outils de dépôt dédiés
  - Moissonne les données d'autres entrepôts de données

Dépôt et recherche dans la suite B2

EUDAT

Dépôt de données

B2DROP
Sync and Exchange Research Data

20 Gb système de gestion de fichier
Permet de définir avec qui échanger les données et de gérer les permissions

B2SHARE
Store and Share Research Data

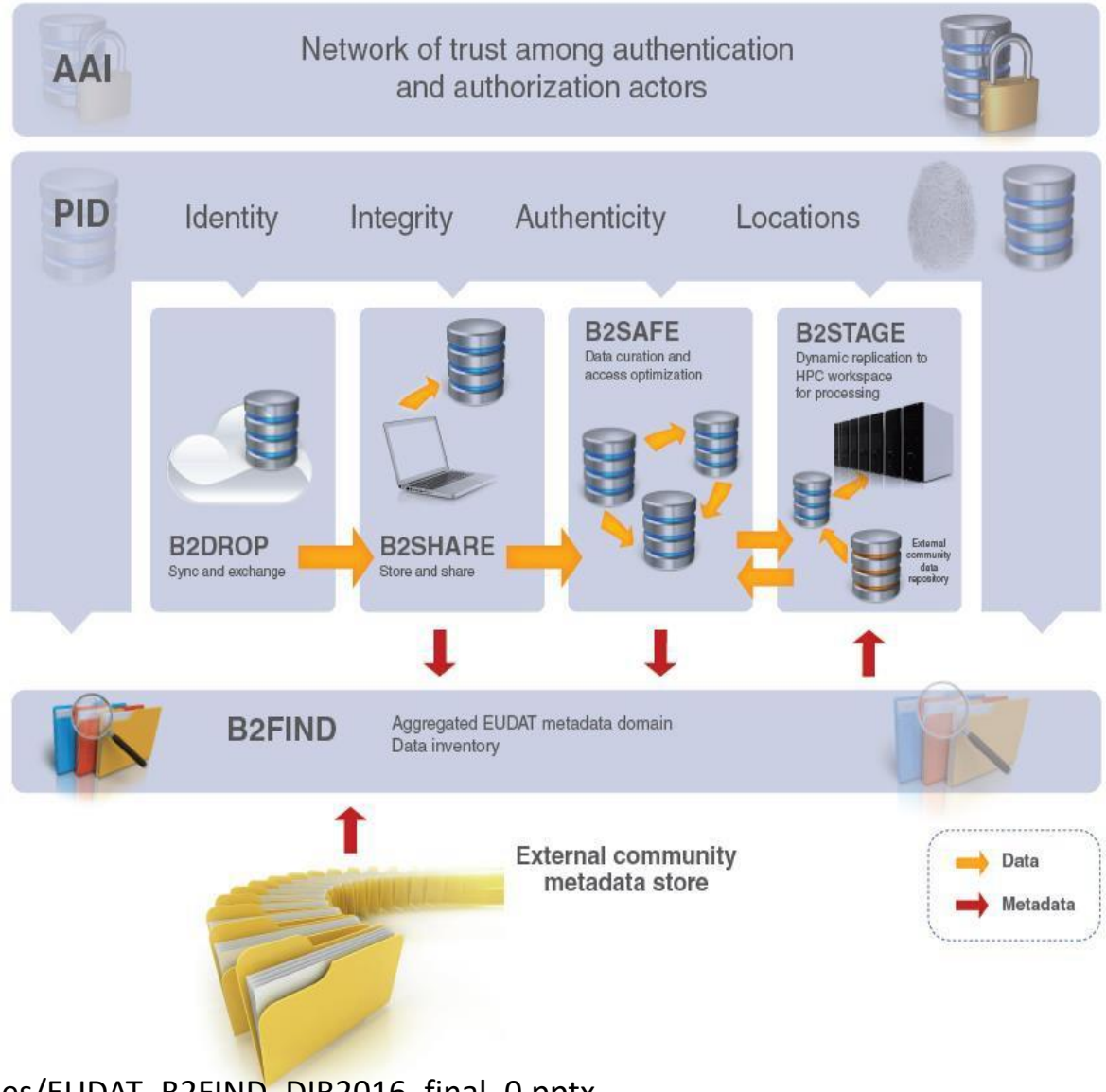
jeux de données volumineux

Archiver les données

B2SAFE
Replicate Research Data Safely

Chercher les données

B2FIND
Find Research Data

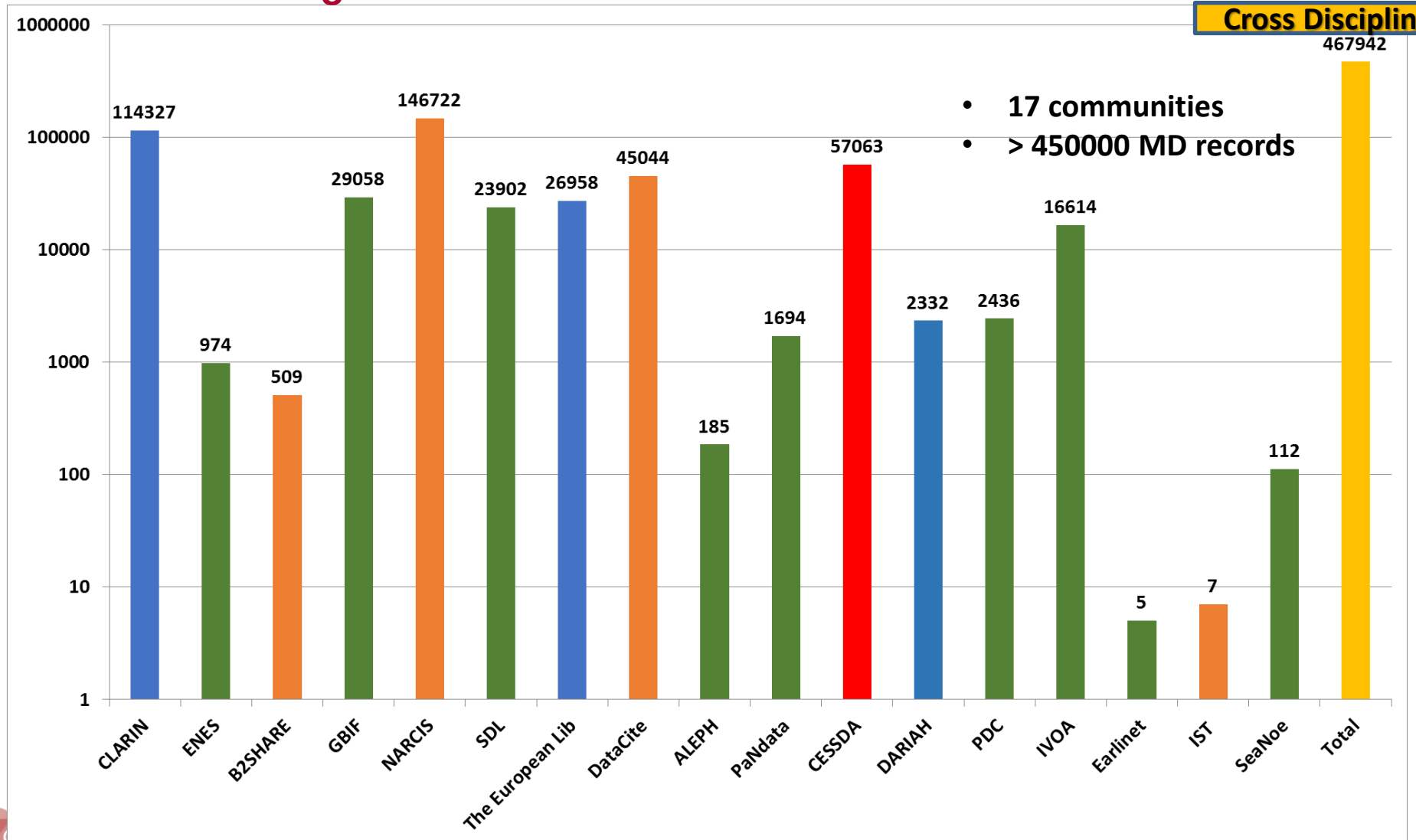


https://www.digitalinfrastructures.eu/sites/default/files/EUDAT_B2FIND_DIR2016_final_0.pptx

B2FIND MD Catalogue

Ingestion status

Humanities
Social Sciences
Natural Sciences
Cross Discipline



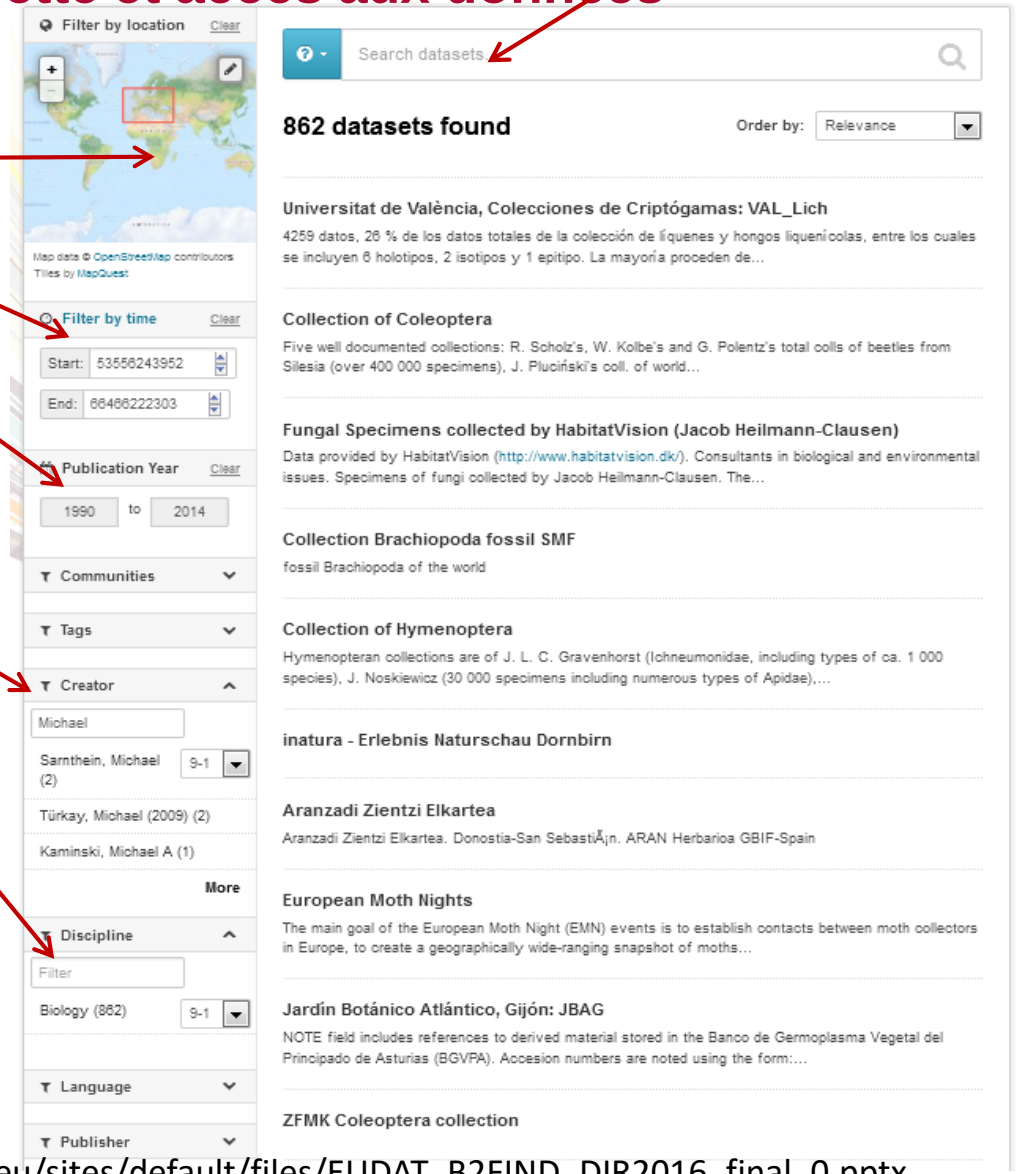
B2FIND Discovery Portal

recherche par facette et accès aux données

Texte libre

Facettes pour :

- Geo spatial
- Couverture temporelle
- Année de publication
- Facettes “texte”
  - Tags
  - Auteur
  - Discipline etc.

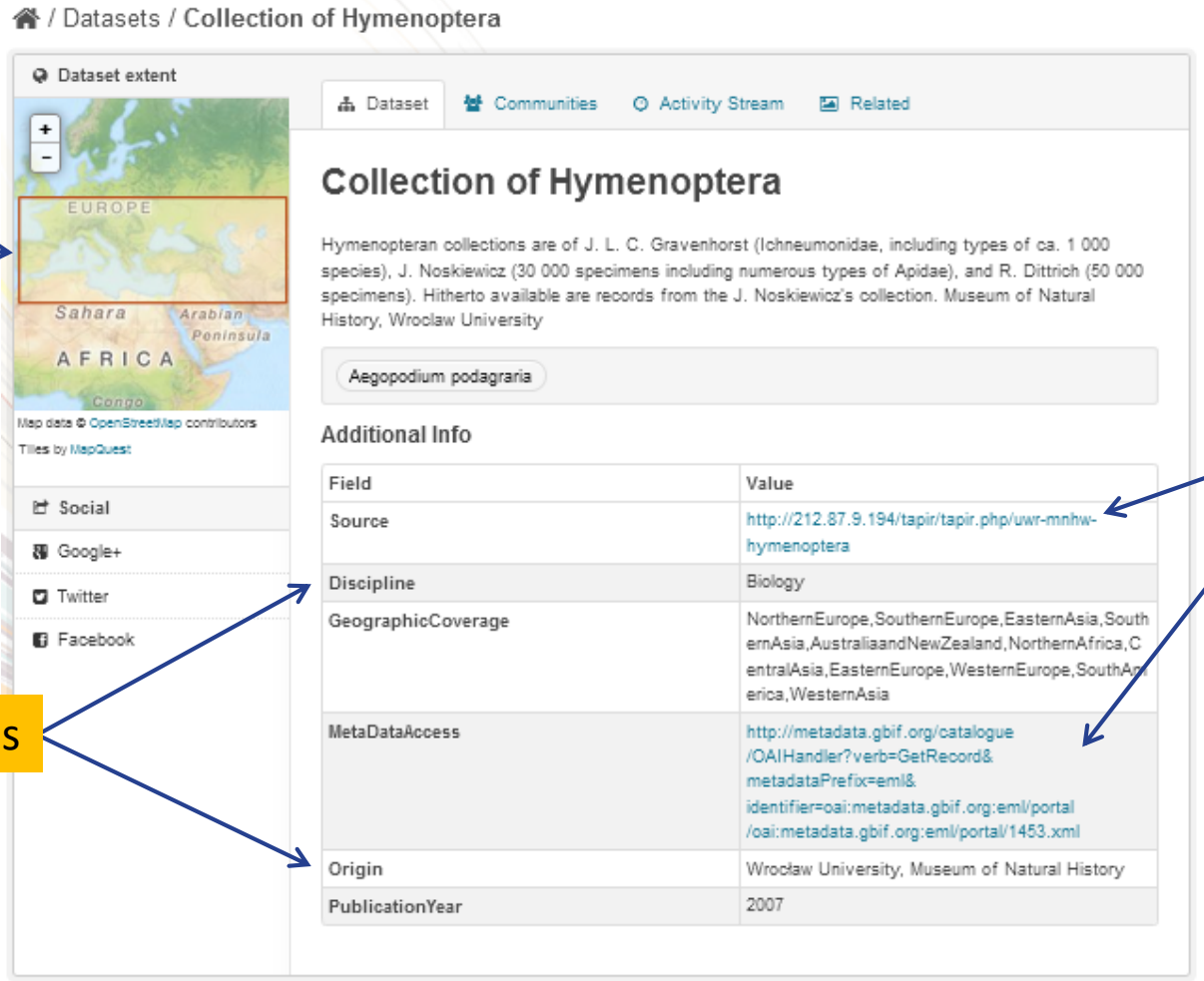


https://www.digitalinfrastructures.eu/sites/default/files/EUDAT_B2FIND_DIR2016_final_0.pptx

Extention spatiale

métadonnées

Lien vers les sources



The Dataverse Project

- Création

Depuis 2006, reprise d'un projet plus ancien. Harvard , Alfred P. Sloan Foundation, National Science Foundation, National Institutes of Health, …

- Contenu

+ de 74 000 datasets de 2171 Dataverse pour 24 installations

- Fonctionnalités intéressantes
  - Outil open source - chaque installation est indépendante. Le Dataverse « Harvard » regroupe tous (ou énormément) de Dataverses de chaque installation
  - Lien entre le journal et le jeux de données
- Modèle de données
  - S'appuie sur DDI (Data Document Initiative)
  - Personnalisable pour chaque Dataverse

DATAVERSE REPOSITORIES - A WORLD VIEW

24 Installations | 2,172 Dataverses | 48,944 Datasets | 2,542,191 Downloads

Dataverse : Outils intégrés

CURRENT INTEGRATIONS | FUTURE INTEGRATIONS

Two Ravens
PIs: James Honaker (Harvard University), Vito D'Orazio (University of Texas at Dallas)

TwoRavens is a system of interlocking statistical tools for data exploration, analysis, and meta-analysis. Read More...

World Map
Collaborators: Boston Area Research Initiative (BARI), The Center for Geographic Analysis (CGA)

WorldMap with Dataverse provides support for geospatial visualization. Read More...

Open Science Framework

The Center for Open Science's Open Science Framework (OSF) is an open source software project that facilitates open collaboration in science research across the lifespan of a scientific project. Read More...

Open Journal Systems

Open Journal Systems (OJS) is a journal management and publishing system that has been developed by the Public Knowledge Project through its federally funded efforts to expand and improve access to research. Read More...

Dataverse package for R on rOpenSci Project
Developed by Thomas Leeper (Department of Government, LSE)

The Dataverse package for R, part of the rOpenSci Project, integrates public data sharing into the reproducible research workflow. Read more...

EZID

EZID (easy-eye-dee) makes it easy to create and manage long-term, globally unique identifiers for your data and sources, ensuring their future discoverability. Read More...

DataCite

DataCite is a leading global non-profit organisation that provides persistent identifiers (DOIs) for research data. Read More...

SHARE

SHARE is building a free, open, data set about research and scholarly activities across their life cycle. Read More...

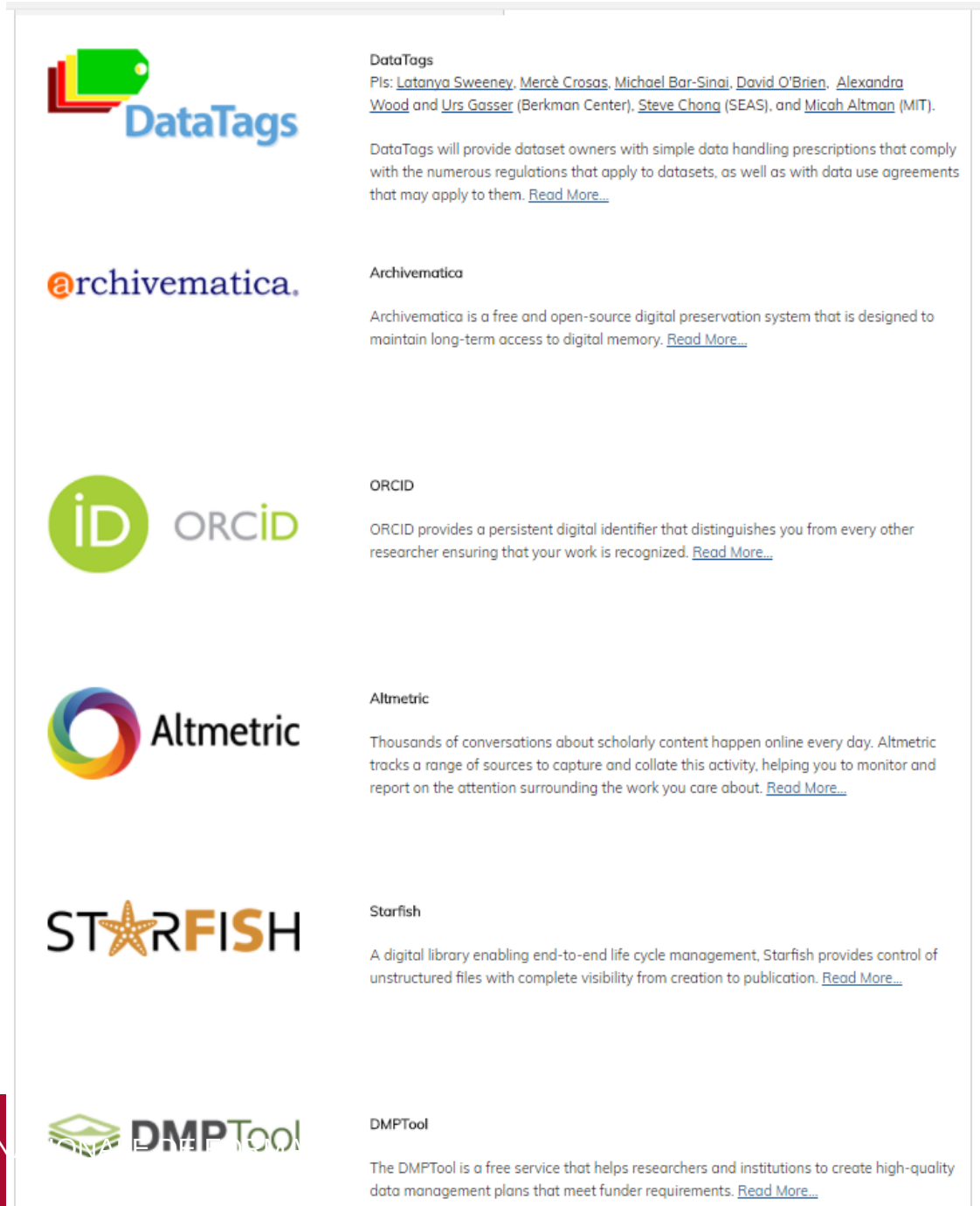
Piwik

Piwik is a free and open source web analytics application that tracks online visits to one or more websites and displays reports on these visits for analysis. Read More...

RSpace

RSpace is an affordable and secure enterprise grade electronic lab notebook (ELN) for researchers to capture and organize data. Read More...

Intégrations futures



DataTags

PIs: Latanya Sweeney, Mercè Crosas, Michael Bar-Sinai, David O'Brien, Alexandra Wood and Urs Gasser (Berkman Center), Steve Chong (SEAS), and Micah Altman (MIT).

DataTags will provide dataset owners with simple data handling prescriptions that comply with the numerous regulations that apply to datasets, as well as with data use agreements that may apply to them. Read More...

Archivematica

Archivematica is a free and open-source digital preservation system that is designed to maintain long-term access to digital memory. Read More...

ORCID

ORCID provides a persistent digital identifier that distinguishes you from every other researcher ensuring that your work is recognized. Read More...

Altmetric

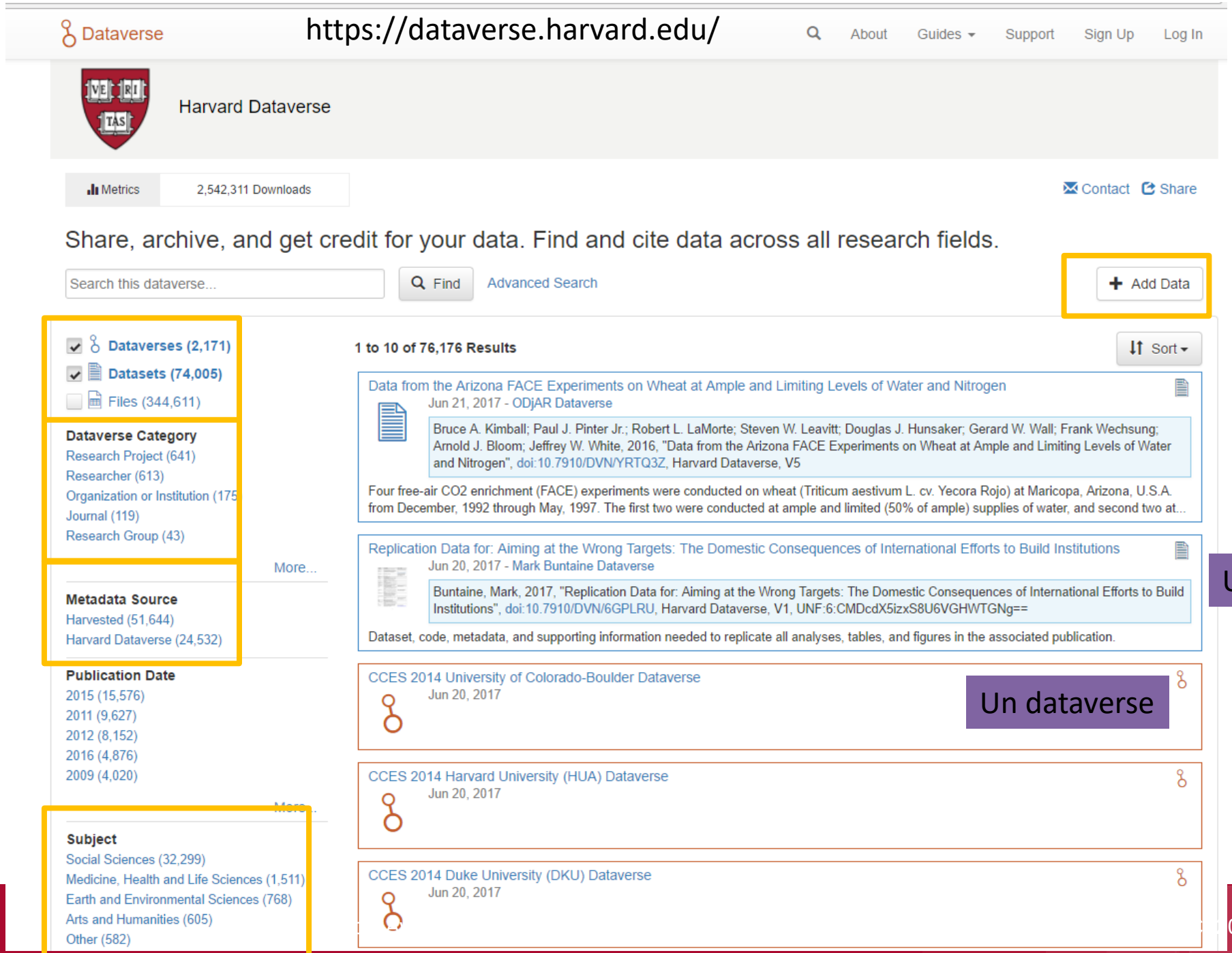
Thousands of conversations about scholarly content happen online every day. Altmetric tracks a range of sources to capture and collate this activity, helping you to monitor and report on the attention surrounding the work you care about. Read More...

Starfish

A digital library enabling end-to-end life cycle management, Starfish provides control of unstructured files with complete visibility from creation to publication. Read More...

DMPTool

The DMPTool is a free service that helps researchers and institutions to create high-quality data management plans that meet funder requirements. Read More...



Add Data
Add Dataverse

Un jeu de données

Un dataverse

Dataverse

About Guides Support Sign Up Log In

ODjAR Dataverse (Open Data Journal for Agricultural Research) www.odjar.org

Harvard Dataverse > ODjAR Dataverse > **Data from the Arizona FACE Experiments on Wheat at Ample and Limiting Levels of Water and Nitrogen**

Metrics 4 Downloads

Un jeu de données

Contact Share

Data from the Arizona FACE Experiments on Wheat at An

Version 5.0

Bruce A. Kimball; Paul J. Pinter Jr.; Robert L. LaMorte; Steven W. Leavitt; Douglas J. Hunsaker; Ge Wechsung; Arnold J. Bloom; Jeffrey W. White, 2016, "Data from the Arizona FACE Experiments on Limiting Levels of Water and Nitrogen", doi:10.7910/DVN/YRTQ3Z, Harvard Dataverse, V5

Description Four free-air CO2 enrichment (FACE) experiments v Arizona, U.S.A. from December, 1992 through May, supplies of water, and second two at ample (350 kg than 50 scientists participated, and they collected a elevated CO2 and its interactions with the water anc including management, soils, weather, physiology, p balance, soil moisture, nitrogen assimilation, and oth

Subject Earth and Environmental Sciences

Keyword Weather, canopy temperature, soil heat flux, ozone, balance, Reflectance and NDVI, Root biomass

Files Metadata Terms Versions

Search this dataset... Find

25 Files

Download

Les données

15826_Creators_Datafiles.docx
MS Word (docx) - 14.7 KB - Jun 13, 2017 - 1 Download
MD5: 5b5544be1f9ffb2f0976782c70d8290b

15826_dataset_description.url
Unknown - 150 bytes - Jun 21, 2017 - 0 Downloads
MD5: 2937348eb4dd9e4080ef1280dd55d3be
Link to the description of the data set in the Open Data Journal for Agricultural Research: http://library.wur.nl/ojs/index.php/ODJAR/article/view/15826

Biomass Yield Area Phenology Management Weather Soil Moisture.ods
application/vnd.oasis.opendocument.spreadsheet - 781.4 KB - Jun 12, 2017 - 0 Downloads
MD5: 5ca916cc0e6177e79b0144229cd8daff
Main spreadsheet with biomass growth, leaf area, growth stage, yield, daily weather, management, soil moisture, and other data generally used for plant growth model validation.

Canopy Leaf Soil Temperatures from Hand-held Infrared Thermometers 1993.ods
application/vnd.oasis.opendocument.spreadsheet - 85.0 KB - Jun 12, 2017 - 0 Downloads
MD5: f73b5d7b9ff552cb51c4520c658a3556
Canopy, soil surface, and leaf temperatures determined near mid-day on many days per season at various view angles with portable hand-held infrared thermometers for 1992-3.

Canopy Leaf Soil Temperatures from Hand-held Infrared Thermometers 1994.ods
application/vnd.oasis.opendocument.spreadsheet - 50.10 KB - Jun 12, 2017 - 0 Downloads
MD5: 8eed3d8136c9c913c69447e864465470
Canopy, soil surface, and leaf temperatures determined near mid-day on many days per season at various view angles with portable hand-held infrared thermometers for 1993-4.

Canopy Leaf Soil Temperatures from Hand-held Infrared Thermometers 1996.ods
application/vnd.oasis.opendocument.spreadsheet - 51.9 KB - Jun 12, 2017 - 0 Downloads
MD5: 69a1326ee34a25e97817d178559558be
Canopy, soil surface, and leaf temperatures determined near mid-day on many days per season at various view angles with portable hand-held infrared thermometers for 1995-6.

Canopy Leaf Soil Temperatures from Hand-held Infrared Thermometers 1997.ods
application/vnd.oasis.opendocument.spreadsheet - 46.7 KB - Jun 12, 2017 - 0 Downloads
MD5: 7d11f27675345c272a4035d214fd9260
Canopy, soil surface, and leaf temperatures determined near mid-day on many days per season at various view angles with portable hand-held infrared thermometers for 1996-7.

fAPAR fraction Absorbed Photosynthetically Active Radiation.ods
application/vnd.oasis.opendocument.spreadsheet - 168.4 KB - Jun 12, 2017 - 0 Downloads
MD5: 82ff17fdc1a37e9f94c485c9ce6338dc
Fraction of absorbed photosynthetically active radiation.

- Création : mai 2O13, Créé par le projet Openaire (UE) et le CERN
- Contenu

tous types de documents , volumétrie

- Fonctionnalités intéressantes
  - Lien direct (signalement et création du rapport)avec le projet Openaire pour le partage des données « Horizon 2020 ». Identification du n° du projet
  - Possibilité de créer des communautés et de définir des droits dans ses communautés.
  - Gestion des accès par dépôt : ouverte, restreinte à une communauté, déposant
  - Embargo possible
  - Créer un DOI pour chaque version du jeu de données
  - API qui permettent d'utiliser Zenodo comme outil de stockage des données
  - Intégration avec Github
  - Altmetrics

- Lien avec Datacite

Recent uploads

June 15, 2017 (v1) Dataset Open Access

View

US Water Network Observed and Inferred Flows

Rising, James

This dataset provides a synthesized national record of river gauges. It includes inferred flows which are based on the structure of the river network and empirical flow relationships. In total, the network contains 22619 gauges (or virtual junction gauges) and over 1 million years of monthly...

Uploaded on June 15, 2017

June 15, 2017 (v7) Software Open Access

View

PCMDI/pcmdi_metrics: PMP 1.1.2

Charles Doutriaux; Paul J. Durack; gleckler1; Zeshawn Shaheen; Jeffrey Painter; Jiwoo Lee; jservonnat; John Krasting

You can install this release via conda either form the environment file conda env create -f pmp-1.1.2.yml or regular conda conda create -n pmp-1.1.2 -c conda-forge -c uvcdat -c pmcdi

Uploaded on June 15, 2017

6 more version(s) exist for this record

June 15, 2017 (v2) Software Open Access

View

stan-dev/math v2.16.0

Daniel Lee; Bob Carpenter; Peter Li; Michael Betancourt; maverickg; Rob Trangucci; Marcus Brubaker; Marco Inacio; Alp Kucukelbir; Mitzi Morris; Krzysztof Sakrejda; seantalts; Charles Margossian; Jeffrey Arnold; bgoodri; wds15; Matt Hoffman; thelseraphim; Avraham Adler; Dustin Tran; Alexey Stukalov; Rob J Goedman; Kevin S. Van Horn; Ben Bales; Juan Sebastián Casallas; Mike Lawrence; Brian Bloniarz; Seth Flaxman; Michael Shvartsman; Amos Waterland

v.2.16.0 (15 June 2017) New Features New append_array function Add categorical_logit_rng function Bug Fixes Align gamma_* function parameter names with documentation Other Update to Eigen 3.3.3 Support g++ 4.9 Fix overload logic in mdivide_left_tri_low so that it calls the var version of...

Uploaded on June 15, 2017

1 more version(s) exist for this record

Zenodo now supports DOI versioning!

Read more about it, in our newest blog post.

Using GitHub?

Just Log in with your GitHub account and click here to start preserving your repositories.

Zenodo in a nutshell

- **Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- **Citeable. Discoverable.** — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- **Communities** — create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
- **Funding** — identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- **Flexible licensing** — because not everything is under Creative Commons.
- **Safe** — your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data.

June 15, 2017

Dataset Open Access

US Water Network Observed and Inferred Flows

Rising, James

This dataset provides a synthesized national record of river gauges. It includes inferred flows which are based on the structure of the river network and empirical flow relationships. In total, the network contains 22619 gauges (or virtual junction gauges) and over 1 million years of monthly data.

The dataset is provide as a saved data file for R, flowdata.RData, containing two variables: nodes and allflow.

`nodes` describes each gauges or virtual junction gauge in the water flow network, representing a combination of gauges from multiple sources. The source is specified by the `collection` column, as follows:

- `rivdis`: From the RivDIS dataset, at http://iridl.ldeo.columbia.edu/SOURCES/.UNH/.CSRC/.RivDIS/index.html?Set-Language=en
- `usgs`: GAGES II river gauge
- `reservoir`: National Inventory of Dams reservoir
- `usgsres`: USGS gauged reservoir
- `canal`: Canal cross county boundaries
- `junction`: Junction node added to capture the structure of rivers from HydroSHEDS

The `colid` specifies the ID of the gauge within its collection, and combined `collection` and `colid` are unique. The location of the gauge is specified by the latitude `lat` and longitude `lon`. The elevation is provided in meters in column `elev`.

The flow data is contained in the `allflow` matrix. This matrix has a column for each node (as many columns as there are rows in `nodes)`, and a row for each month starting from Jan. 1915. All the flows are in units of m^3 / s.

Files (68.7 MB)

Name	Size	
flowdata.RData md5:00111afd2fa7c3e704d184cbb18d3782	68.7 MB	Download

Indexed in OpenAIRE

Publication date:
June 15, 2017

DOI:
DOI 10.5281/zenodo.809469

Keyword(s):
streamflow, gap-filling, gauges

Communities:
Zenodo

License (for files):
Creative Commons Attribution 4.0

Versions

Version 1 10.5281/zenodo.809469 Jun 15, 2017

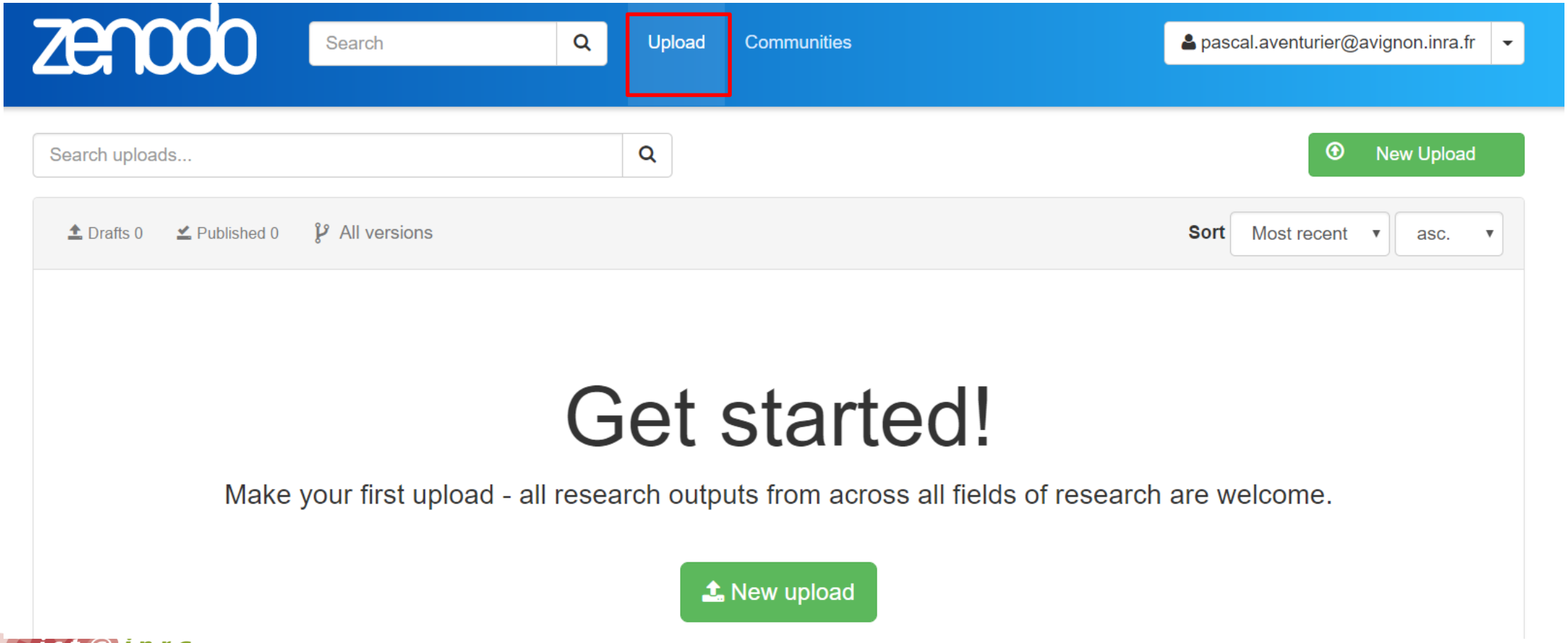
Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.809468. This DOI represents all versions, and will always resolve to the latest one. Read more.

Share

Cite as

Rising, James. (2017). US Water Network Observed and Inferred Flows [Data set]. Zenodo. http://doi.org/10.5281/zenodo.809469

Dépôt dans Zenodo



New upload

Instructions: (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Files

Choose files | Start upload

Filename (1 files)	Size	Progress	Delete
AGRISEMANTICS19JUIN17.xls md5:08aef6d02cb6b96b54e74f54e54fd1c4	6 Kb	✔	

Note: File addition, removal or modification are not allowed after you have published your upload. This is because a Digital Object Identifier (DOI) is registered with DataCite for each upload.

(minimum 1 file required, max 50 GB per dataset - contact us for larger datasets)

Upload type — required

Publication | Poster | Presentation | **Dataset** | Image | Video/Audio | Software | Lesson

Basic information — required

Digital Object Identifier

e.g. 10.1234/foo.bar

Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

Pre-reserve DOI

Publication date *

2017-06-27

Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

Title *

dataset for the agrisemantic bibliometric analysis

Required.

Authors *

Aventurier, Pascal | INRA

+ Add another author | Affiliation

Type de
De données

Métadonnées 1

Description *

Source

ceci est une description

Required.

Keywords

+ Add another keyword

Additional notes

Optional.

License required

Access right *

- Open Access
- Embargoed Access
- Restricted Access
- Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

Conditions

Source

WORKING GROUP

Métadonnées
Licences

Communities recommended

Any user can create a community collection on Zenodo (browse communities). Specify communities which you wish your upload to appear in. The owner of the community will be notified, and can either accept or reject your request.

Communities INRA Datasets

+ Add another community

Funding recommended

Zenodo is integrated into reporting lines for research funded by the European Commission via OpenAIRE (http://www.openaire.eu). Specify grants which have funded your research, and we will let your funding agency know!

Grants Start typing a grant number, name or abbreviation...

Optional. European Commission FP7 and Horizon 2020 grants only. For general funding acknowledgements, please use the *Additional Notes* field.
Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in OpenAIRE.

+ Add another grant

Related/alternate identifiers recommended

Contributors optional

References optional

Journal optional

Conference optional

Book/Report/Chapter optional

Thesis optional

Subjects optional

Delete

Métadonnées Recommandées/ optionnelles

Jeu de données final

Choisir un entrepôt

Sylvie Cocaud

Les entrepôts de données de recherche

ist@inra
INRA
SCIENCE & IMPACT

Choisir un entrepôt : prendre en compte les recommandations…

- **de la discipline dont relève la recherche**
  - Genbank, UniProtKB, Protein Data Bank in Europe (PDBe)…etc
- **des tutelles**
  - Inra : code sources : forge logicielle SourceSup ; données n'ayant pas vocation à aller dans un entrepôt disciplinaire : communauté Inra Zenodo puis futur portail datainra.fr
- **des financeurs**
  - Wellcome Trust Data repositories (13 entrepôts recommandés)
  - H2020 : **entrepôt garantissant la gratuité de l'accès**, de l'extraction, de l'exploitation, de la reproduction et de la dissémination.
- **des revues pour les données qui seront publiées**
  - Recommandent de + en + le dépôt dans un entrepôt disciplinaire (ou à défaut généraliste), voire intègrent un entrepôt dans le processus de publication (Nature + Figshare, Open Journal System + Dataverse…)
    - Nature : +90 entrepôts recommandés
    - CellPress : liste par type de données

Choix d'un entrepôt : les critères qui entrent en jeu dans la publication des données de recherche

Formats acceptés
- Formats des fichiers
- Organisation du contenu
- Prévisualisation selon formats
- Limites de volumétrie

Documentation des données
- Minimum de métadonnées standards (DC, Datacite), méta de provenance
- Métadonnées de domaine riches
- Description et métadonnées au niveau du jeu + au niveau file

Licences
- Terms of use pour le dépôt
- Choix des licences
- Embargo, accès contrôlé

Coûts de la publication et de l'accès
- Garantir accès libre

Validation des données
- Avant publication (checksum, contrôle méta obligatoires)
- Worflow de curation
- Après publication (stats d'usage, popularité)

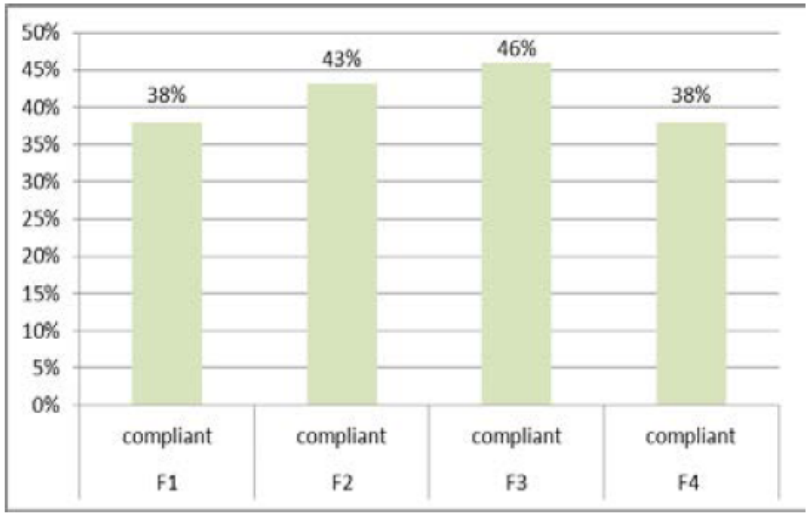
Disponibilité présente et future des données
- Archivage pérenne
- Certification de l'entrepôt
- Ressources, financement pérennes

Découverte et accès
- Recherche (keywords, navigation, facettes, localisation, proximité…)
- Recommandations
- OAI-PMH
- API
- Linked data
- Visibilité de l'entrepôt, Intégration dispositifs [inter]nationaux
- Optimisation pour les moteurs de recherche

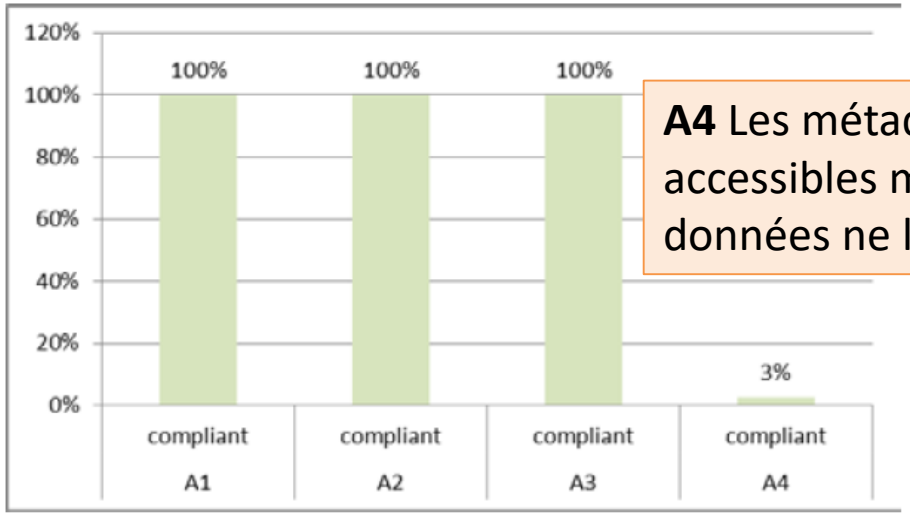
Citation
- Identification unique pérenne (données, chercheurs)
- Affichage, export dans ≠ formats, intégration dans site Web
- Partage sur les réseaux sociaux
- Citation au niveau version
- Lien avec publi, avec datasets
- Génération datapaper, Intégration dans processus de publication

F1 Les données et les métadonnées sont identifiées par un identifiant global unique et pérenne

FINDABLE



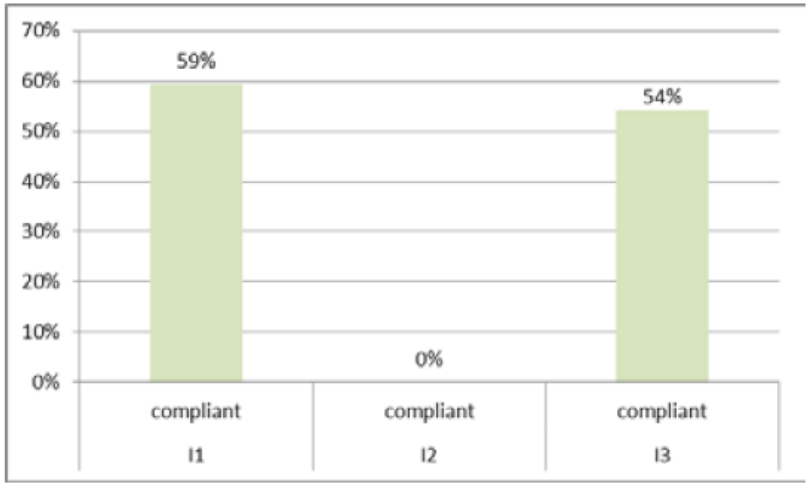
ACCESSIBLE



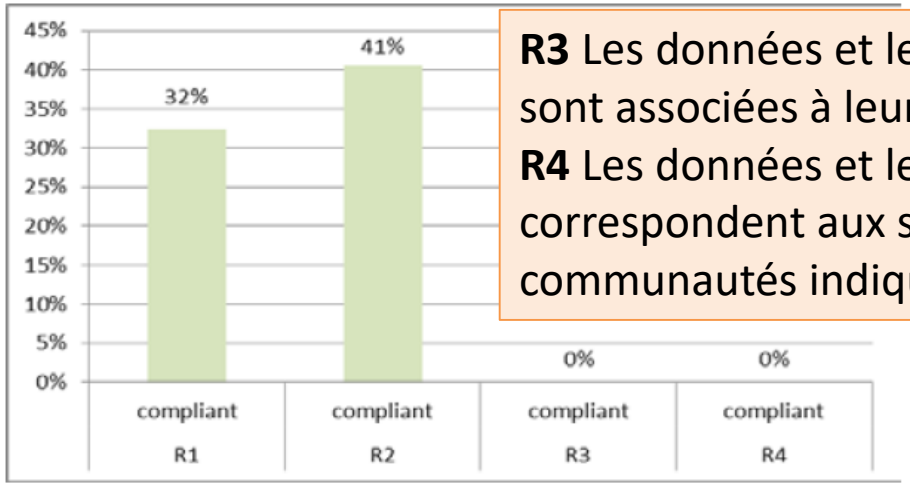
A4 Les métadonnées sont accessibles même quand les données ne le sont plus

I2 Les données et les métadonnées utilisent des vocabulaires qui respectent les principes FAIR

INTEROPERABLE



RE-USABLE



R3 Les données et les métadonnées sont associées à leur provenance.
R4 Les données et les métadonnées correspondent aux standards des communautés indiquées.

(Dunning, de Smaele, & Böhmer, 2017) 37 entrepôts

Traduit de (Hagstrom, 2014)

Les principes FAIR

Faciles à trouver

- F1 Les données et les métadonnées sont identifiées par un identifiant global unique et pérenne.
- F2 Les métadonnées décrivant les données sont riches.
- F3 Les données et les métadonnées sont enregistrées et indexées dans un dispositif permettant de les rechercher.
- F4 Les métadonnées spécifient l'identifiant de la donnée.F1 (meta)data are assigned a globally unique and eternally persistent identifier

Accessibles

- A1 Les données et les métadonnées sont accessibles par leur identifiant via un protocole de communication standardisé.
- A1.1 Le protocole utilisé est ouvert, libre et peut être implémenté de manière universelle.
- A1.2 Le protocole utilisé permet l'authentification et l'autorisation si besoin.
- A2 Les métadonnées sont accessibles même quand les données ne le sont plus

Interopérables

- I1 Les données et les métadonnées utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances.
- I2 Les données et les métadonnées utilisent des vocabulaires qui respectent les principes FAIR.
- I3 Les données et les métadonnées incluent des liens vers d'autres (méta)données

Ré-utilisables

- R1 Les données et les métadonnées ont des attributs multiples et pertinents.
- R1.1 Les données et les métadonnées sont mises à disposition selon une licence explicite et accessible.
- R1.2 Les données et les métadonnées sont associées à leur provenance.
- R1.3 Les données et les métadonnées correspondent aux standards des communautés indiquées.

Choisir un entrepôt : les répertoires d'entrepôts

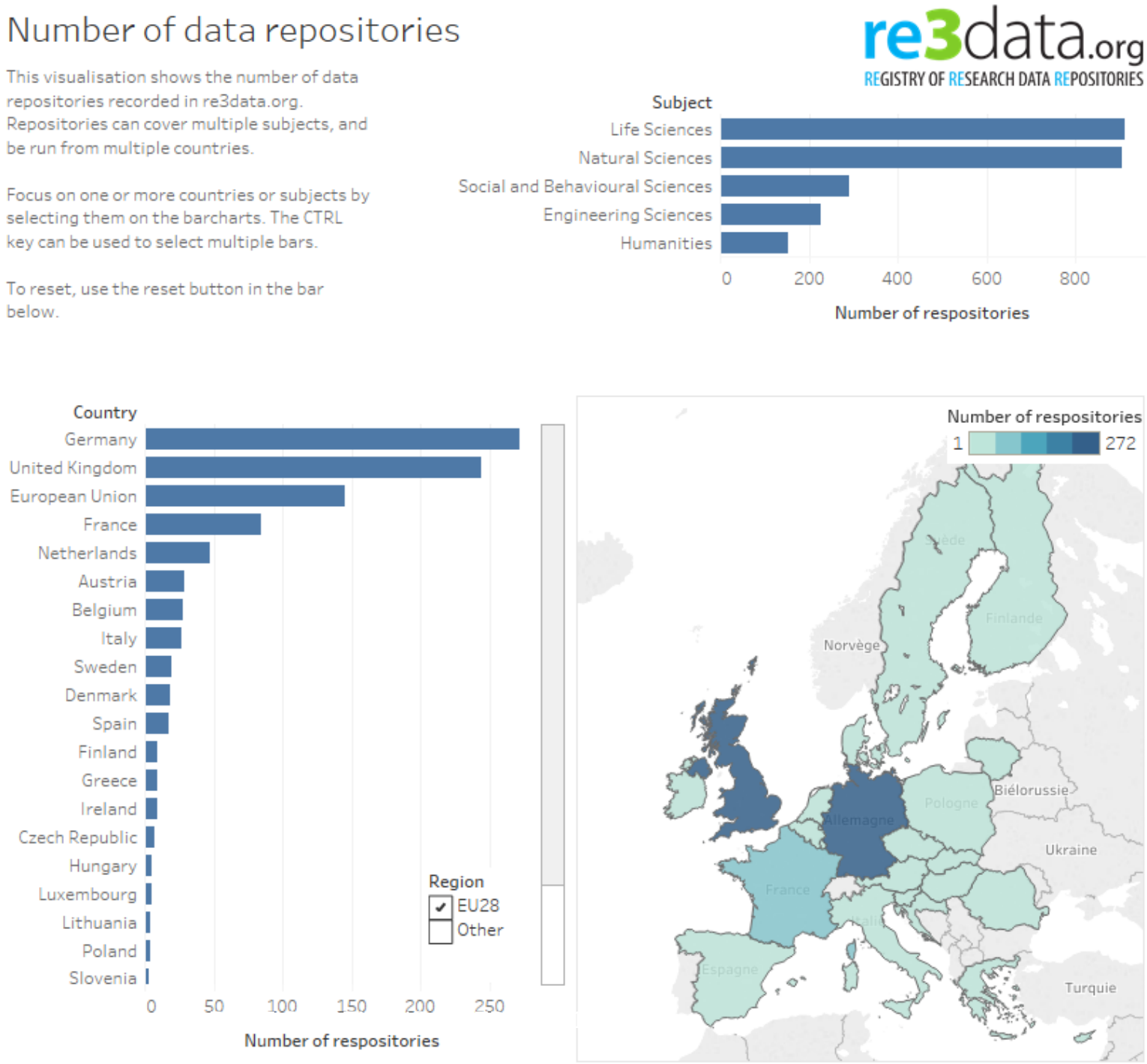
- re3data.org (1859 entrepôts répertoriés) tous domaines
- bioSharing (917 bases de données répertoriées) domaines des sciences de la vie, de l'environnement, de la biomédecine
- OpenDOAR (178 entrepôts de données sur 3344 répertoriés)
- OpenAIRE Locate data provider

Chiffres d'après les consultation des sites entre mai et juin 2017

Choix d'un entrepôt : le répertoire re3data

It is part of DataCite

Updates are provided by re3data.org and through community contributions (…) it is not fully complete



Choix d'un entrepôt : re3data

- Accès aux données : open / restricted / closed
- Conditions d'usage et licenses liées aux données fournies par l'entrepôt
- L'entrepôt décrit sa politique
- Identification des jeux de données par un identifiant persistant, unique et citable (DOI, URN, ARK, Handle, Purl)
- Entrepôt certifié ou utilisant un standard

Repository details

NSF Arctic Data Center

General | Institutions | Terms | Standards

Name of repository	NSF Arctic Data Center
Repository URL	https://arcticdata.io
Subject(s)	Plant Sciences; Zoology; Chemistry; Physics; Soil Sciences; Basic Forest Research; Oceanography; Atmospheric Science; Geochemistry, Mineralogy and Crystallography; Geography; Water Research; Natural Sciences; Humanities and Social Sciences; Biology; Life Sciences; Agriculture, Forestry, Horticulture and Veterinary Medicine; Agriculture, Forestry, Horticulture and Veterinary Medicine; Atmospheric Science and Oceanography; Geosciences (including Geography)
Description	The Arctic Data Center is the primary data and software repository for the Arctic section of NSF Polar Programs. The Center helps the research community to reproducibly preserve and discover all products of NSF-funded research in the Arctic, including data, metadata, software, documents, and provenance that links these together. The repository is open to contributions from NSF Arctic investigators, and data are released under an open license (CC-BY, CC0, depending on the choice of the contributor). All science, engineering, and education research supported by the NSF Arctic research program are included, such as Natural Sciences (Geoscience, Earth Science, Oceanography, Ecology, Atmospheric Science, Biology, etc.) and Social Sciences (Archeology, Anthropology, Social Science, etc.). Key to the initiative is the partnership between NCEAS at UC Santa Barbara, DataONE, and NOAA's NCEI, each of which bring critical capabilities to the Center. Infrastructure from the successful NSF-sponsored DataONE federation of data repositories enables data replication to NCEI, providing both offsite and institutional diversity that are critical to long term preservation.
Contact	support@arcticdata.io info@arcticdata.io https://arcticdata.io/team/
Content type(s)	Databases; Networkbased data; Images; Structured graphics; Audiovisual data; Scientific and statistical data formats; Raw data; Plain text; Structured text; Archived data; Source code; Standard office documents; Software applications
Keyword(s)	Arctic
Repository size	The Arctic Data Center holds over 4.8TB of data in > 3800 data sets comprised of over 480,000 data files.
Repository type(s)	disciplinary

Repository details

NSF Arctic Data Center

General | Institutions | Terms | Standards

Persistent identifier system(s)	DOI
Name of the repository software	other
Versioning	yes
Author identifier system(s)	ORCID
Enhanced Publication	unknown
Quality management	yes

Application programming interfaces (2)

API type	REST
URL	https://arcticdata.io/catalog/#api
API type	NetCDF
URL	https://arcticdata.io/submit/#license

Metadata standards (6)

Metadata standard name	EML - Ecological Metadata Language
Metadata standard scheme	DCC
Metadata standard URL	http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-languag
Metadata standard name	ISO 19115
Metadata standard scheme	DCC
Metadata standard URL	http://www.dcc.ac.uk/resources/metadata-standards/iso-19115
Metadata standard name	CF (Climate and Forecast) Metadata Conventions
Metadata standard scheme	DCC
Metadata standard URL	http://www.dcc.ac.uk/resources/metadata-standards/cf-climate-and-forecast-metadata
Metadata standard name	PROV
Metadata standard scheme	DCC
Metadata standard URL	http://www.dcc.ac.uk/resources/metadata-standards/prov
Metadata standard name	FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata
Metadata standard scheme	DCC
Metadata standard URL	http://www.dcc.ac.uk/resources/metadata-standards/fgdccsdgm-federal-geographic-data-committee-content-standard-digital-ge
Metadata standard name	DataCite Metadata Schema
Metadata standard scheme	DCC
Metadata standard URL	http://www.dcc.ac.uk/resources/metadata-standards/datacite-metadata-schema

Choix d'un entrepôt : la certification

- **Le réseau européen "European Framework for Audit and Certification of Digital Repositories" créé en 2010 constitue un cadre pour l'audit et la certification des entrepôts numériques.**
  3 niveaux de certification :

1. **certification de base** accordée aux entrepôts ayant obtenu le Data Seal of Approval (procédure d'**autoévaluation**).
2. **certification étendue** basée sur les normes ISO 16363 ou DIN 31644 en Allemagne (**autoévaluation** validée par un tiers) ;
3. **certification formelle** ISO 16363 ou DIN 31644 réalisée par des **experts accrédités**.

Place de l'entrepôt dans le cycle du projet de recherche

Avant le projet

- Réfléchir aux données produites, à leur diffusion et aux **entrepôts** concernés, aux coûts de préservation des données
- Commencer à rédiger un plan de gestion des données

Pendant le projet

- Stocker les données dans différents supports et lieux sécurisés
- Déposer dans un **entrepôt** les données à partager avec des tiers identifiés (si peer review nécessaire par ex.)

A la fin du projet

- Déposer en open access, dans un **entrepôt**, toutes les données qui peuvent l'être
- Déposer les données à archiver sur une plateforme d'archivage

DMP OPIDoR

zenodo

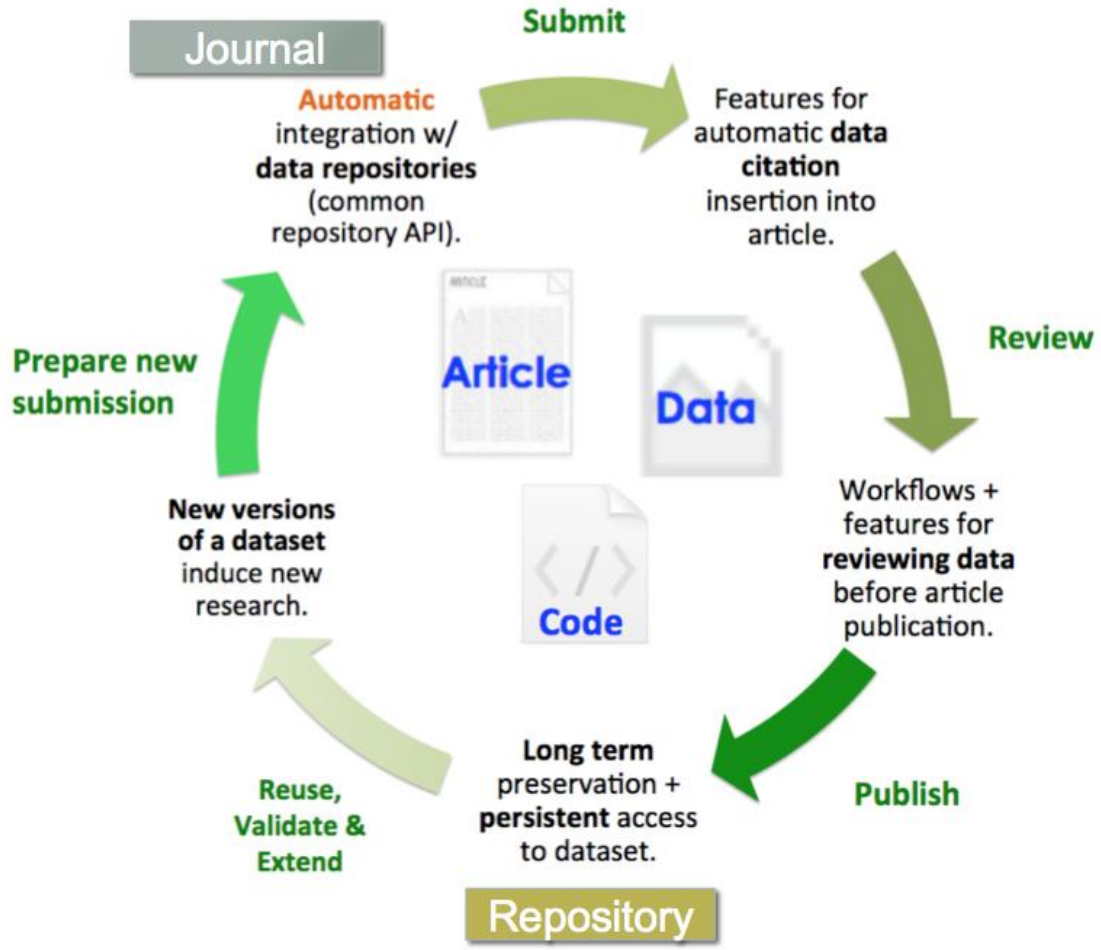
DRYAD

PANGAEA.

INES
Centre Informatique National de l'Enseignement Supérieur

ist@inra

Intégration entrepôt – workflow de publication



Lifecycle of an automated and integrated journal and data publishing workflow (Altman et al., 2015)

[Open Journal Systems](#)
+ Dataverse

[data2paper](#)

[Database Linking Tool](#) (Elsevier)

[Data Literature Interlinking Service](#)

En conclusion

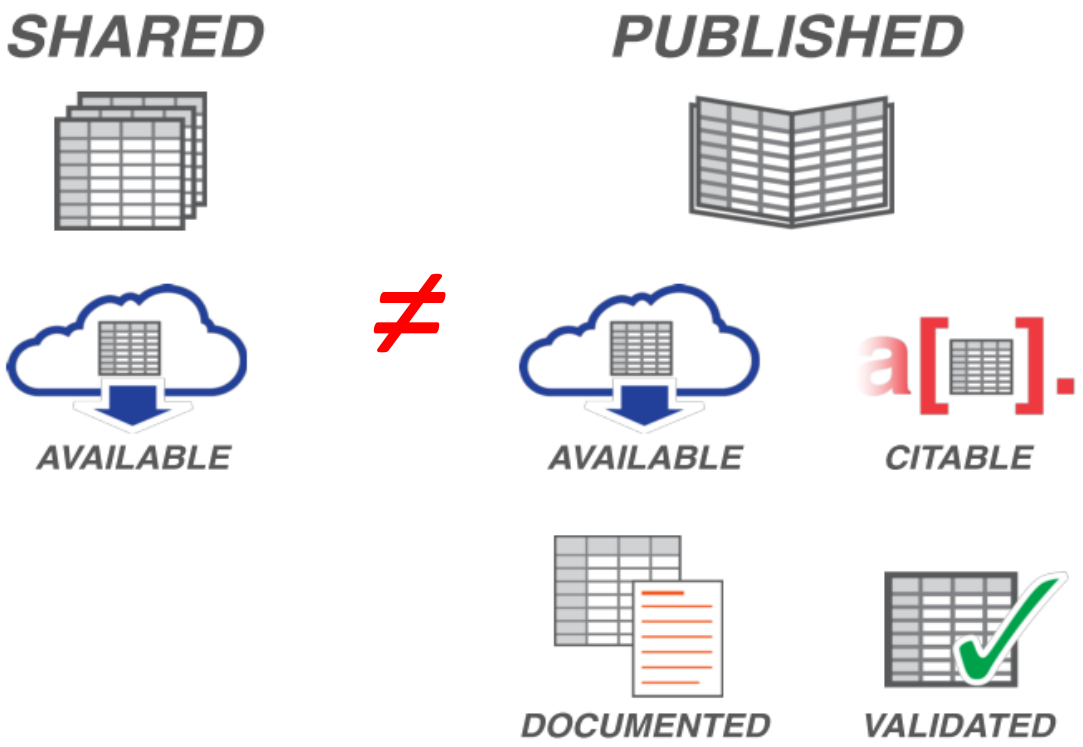
Sylvie Cocaud

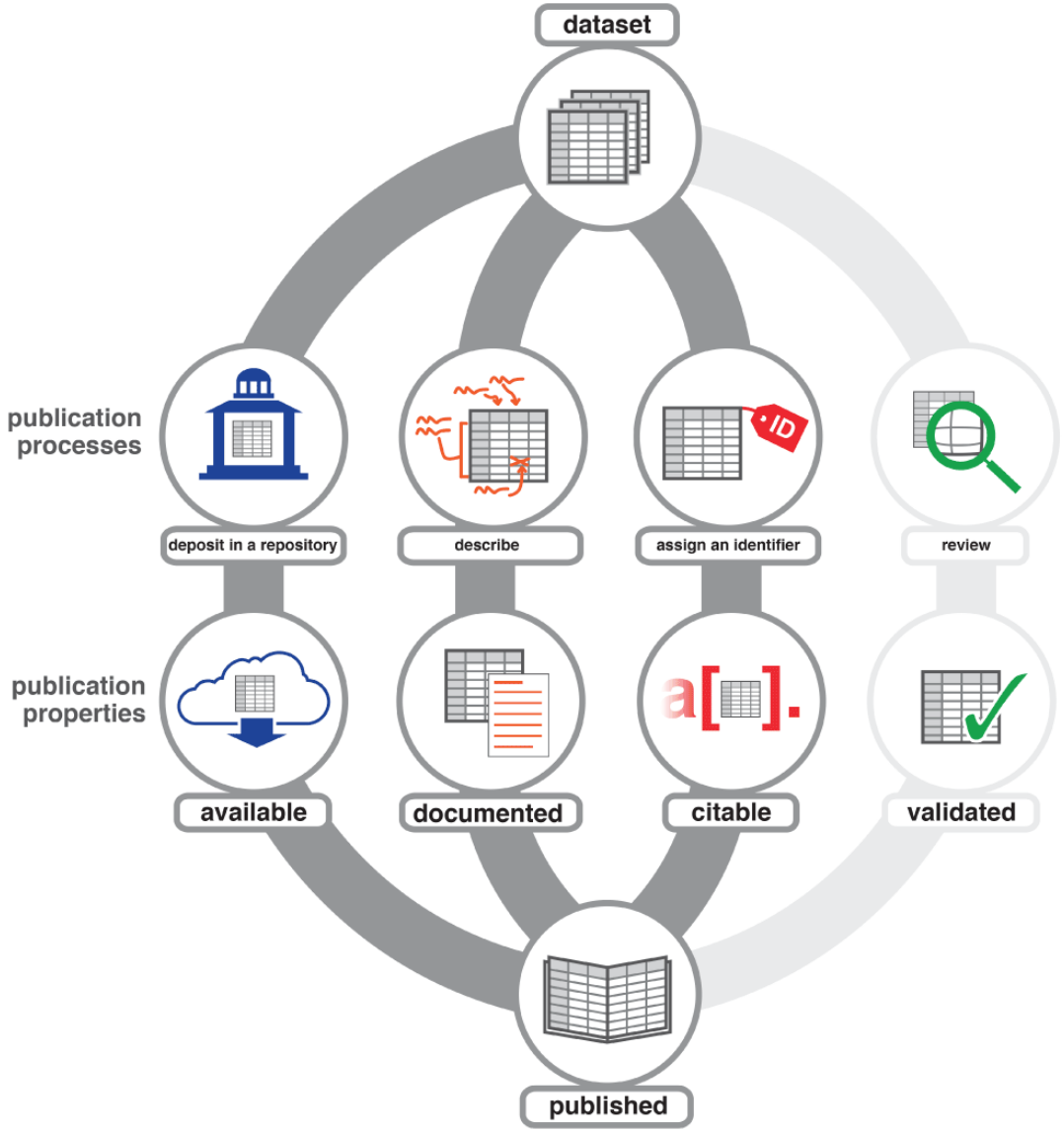
Les entrepôts de données de recherche

ist@inra

INRA
SCIENCE & IMPACT

Partager ses données vs publier ses données



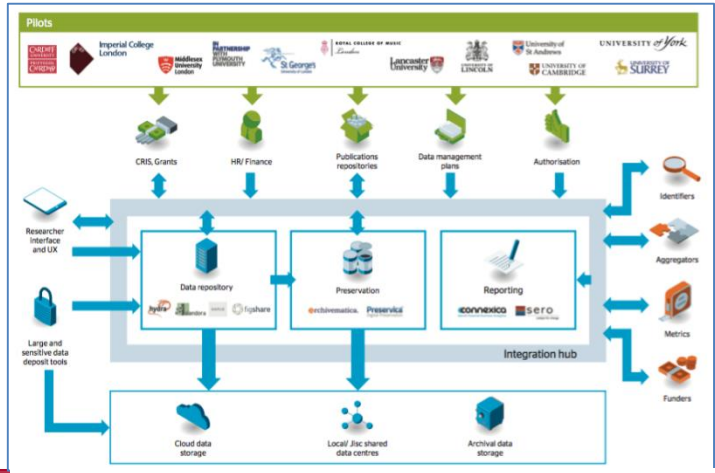


(Kratz & Strasser, 2014)

Déposer ses données dans un entrepôt facilite la découverte et la réutilisation des données

- **Les entrepôts sont scannés par des outils de recherche spécifiques**
  - Datacite search
  - Data Citation Index (Thomson Reuters)
  - Data search (Elsevier)
  - …

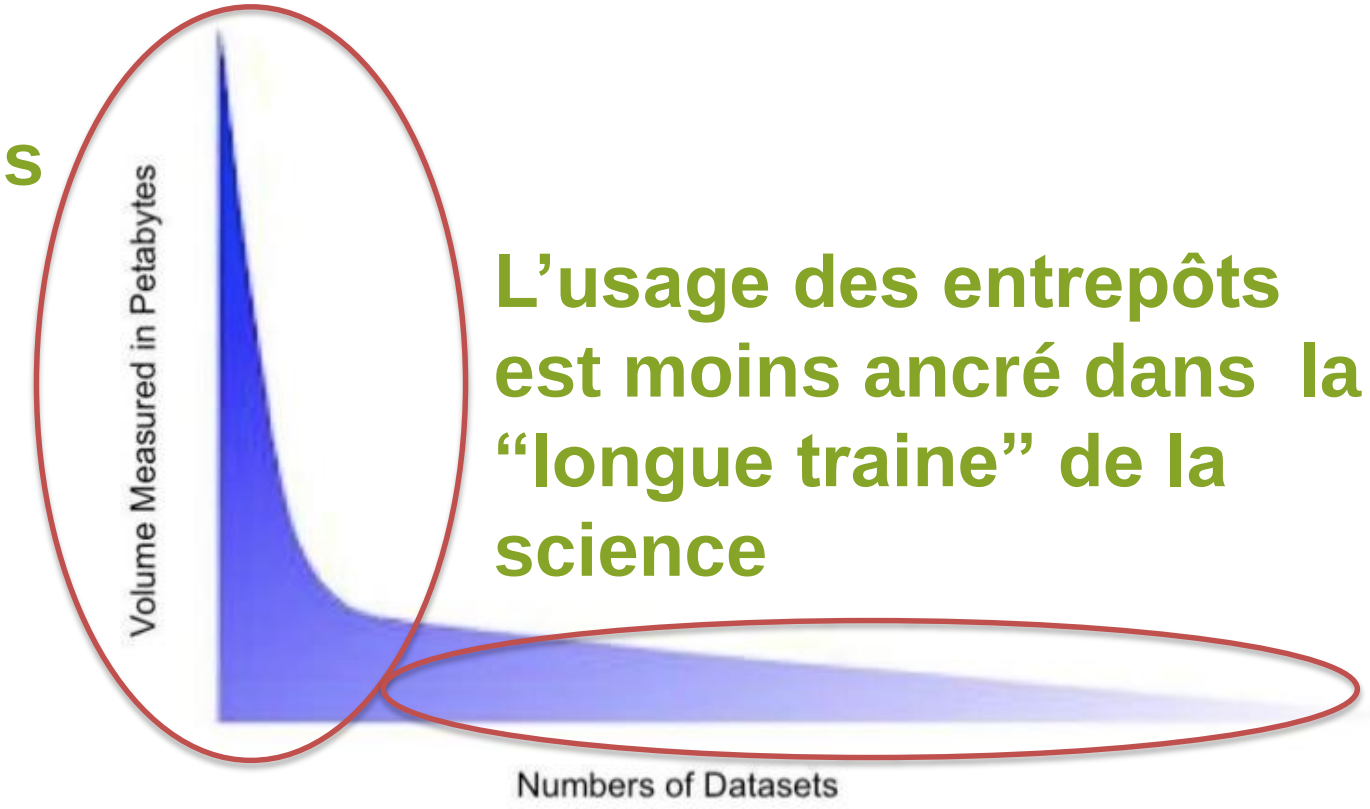
- **et moissonnés par des catalogues, intégrateurs, infrastructures européennes de données… de plus en plus nombreux**



Research Data Shared Service (JISC)

Utilisation des entrepôts par les scientifiques

- **Les entrepôts sont largement et depuis longtemps utilisés dans les domaines qui produisent de grands volumes de données :**
  - astronomie, physique des particules,
  - génétique, sciences de l'environnement.



L'usage des entrepôts est moins ancré dans la "longue traine" de la science

Humphrey (2014)

Références

- **Altman, M., Castro, E., Crosas, M., Durbin, P., Garnett, A., & Whitney, J.** (2015). Open Journal Systems and Dataverse Integration– Helping Journals to Upgrade Data Publication for Reusable Research. *The Code4Lib Journal*, (30). Consulté à l'adresse http://journal.code4lib.org/articles/10989
- **Amitrano, D., Lejeune, S., & Hobléa, F.** (2017, mars). Données d'observation des falaises, comment partager des données hétérogènes. Les cas du St Eynard et du Mont Granier | UMS GRICAD. Présenté à Datalabo 2017. Gérer les données de la recherche : partage de bonnes pratiques, Grenoble. Consulté à l'adresse https://datalabo2017.sciencesconf.org/142248
- **Castro, E.** (2016, 25/02). Dataverse: Helping Researchers Publish Their Data Through Automation. Data & Analytics présenté à 11th International Digital Curation Conference. Visible data, invisible infrastructure, Amsterdam. Consulté à l'adresse https://www.slideshare.net/EleniCastro/dataverse-helping-researchers-publish-their-data-through-automation
- **Dojat, M.** (2017, mars). FLI-IAM : Une infrastructure nationale pour la gestion des données et des outils en imagerie in vivo | UMS GRICAD. Présenté à Gérer les données de la recherche: partage de bonnes pratiques - Sciencesconf.org, Grenoble. Consulté à l'adresse https://gricad.univ-grenoble-alpes.fr/video/fli-iam-infrastructure-nationale-gestion-donnees-et-outils-en-imagerie-vivo
- **Dunning, A., de Smaele, M., & Böhmer, J.** (2017, 23/02). Are the FAIR Data Principles fair? Présenté à IDCC2017. 12th International Digital Curation Conference. Upstream, Downstream: embedding digital curation workflows for data science, scholarship and society, Edinburgh. Consulté à l'adresse http://www.dcc.ac.uk/webfm_send/2480
- **Hagstrom, S.** (2014, septembre 3). The FAIR Data Principles. Consulté 30 juin 2017, à l'adresse https://www.force11.org/group/fairgroup/fairprinciples
- **Jones, S.** (2017). EUDAT, OpenAIRE & DMPs. https://doi.org/10.23728/b2share.49c690be738743aebec8be39413bd672
- **Kratz, J., & Strasser, C.** (2014). Data publication consensus and controversies. F1000Research. https://doi.org/10.12688/f1000research.3979.1
- **OCDE, Caruso, J., Nicol, A., & Archambault, É.** (2007). Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics (p. 29). Consulté à l'adresse http://recolecta.fecyt.es/sites/default/files/contenido/documentos/Third%20report_OA_Data_Policies.pdf
- **Office québécois de la langue française.** (2002). métadonnée. Consulté 30 juin 2017, à l'adresse http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8869869
- **Pomart, J.** (2014, juillet 18). AAF / Section Aurore: Un groupe de travail sur les données de la recherche [Billet]. Consulté 30 juin 2017, à l'adresse http://archivesfmsh.hypotheses.org/1209
- **Thessen, A. E., & Patterson, D. J.** (2011). Data issues in the life sciences. ZooKeys, (150), 15-51. https://doi.org/10.3897/zookeys.150.1766

Sites

- biosharing https://biosharing.org/
- Data Seal of Approval https://www.datasealofapproval.org/
- DataCite search https://search.datacite.org/
- Data Citation Index http://wokinfo.com/products_tools/multidisciplinary/dci/
- Data Literature Interlinking Service https://dliservice.research-infrastructures.eu/#/
- Data2paper https://www.data2paper.org/
- DataSearch https://datasearch.elsevier.com/#/
- Dataverse https://dataverse.harvard.edu/
- Disciplinary Metadata (DCC) http://www.dcc.ac.uk/resources/metadata-standards
- Eudat https://www.eudat.eu/
- FACILE - Service de validation de formats (CINES) https://facile.cines.fr/
- Licence Ouverte etalab version 2.0 https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf
- Nakala https://www.nakala.fr/
- NSF Arctic Data Center https://arcticdata.io/
- Open Data Commons Open Database License (ODbL) https://opendatacommons.org/licenses/odbl/
- OpenAIRE https://www.openaire.eu/
- OpenDOAR http://www.opendoar.org/
- re3data http://www.re3data.org/
- European Framework for Audit and Certification of Digital Repositories TrustedDigitalRepository.eu
- Sustainability of Digital Formats: Planning for Library of Congress Collections https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml
- Zenodo https://zenodo.org/

Merci pour votre attention

sylvie.cocaud@inra.fr et pascal.aventurier@inra.fr
INRA (DIST)