

Open Sciences et FAIR in Design



Céline QUENTIN
Mediterranean Institute of Oceanography
Aix Marseille Univ - CNRS - IRD – Univ. Toulon

☐ SOMMAIRE

1. Open Science : pourquoi FAIRe ?
2. Le cycle de vie de la donnée
3. Méta données et vocabulaire contrôlé
4. OGC SensorThings
5. Outils complémentaires



1. Open Science : pourquoi FAIRe ?

1. Open Science : pourquoi FAIRe?

L'Open Science est un mouvement visant à rendre
les résultats, méthodes et données scientifiques
accessibles à tous.

Elle repose sur **la transparence, le partage et la reproductibilité.**



La vision de l'**European Open Science Cloud** est de mettre en place un système en Europe pour **trouver** et **accéder** aux **données** et aux **services** pour la **recherche et l'innovation**. Il s'agit d'aider les chercheurs à **stocker, partager, traiter, analyser** et **réutiliser** les données **FAIR** de la recherche au sein et **au-delà des disciplines et des frontières**. La connexion entre les entrepôts de données et les services sera déterminante pour que la science ouverte progresse en Europe.

Enjeux :

- Favoriser la collaboration scientifique internationale
- Accélérer l'innovation
- Renforcer la confiance dans la recherche
- Valoriser les données financées par l'argent public

Exemples de mise en œuvre :

HAL, Software Heritage, Zenodo, arXiv, ...

Ressources utiles :

- Open Science – European Commission :
https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en
- Science Ouverte – Ministère français :
<https://www.enseignementsup-recherche.gouv.fr/fr/la-science-ouverte-46586>

1. Open Science : pourquoi FAIRe?

Plan national pour la Science Ouverte

- les résultats de la recherche scientifique ouverts à tous, sans entrave, sans délai, sans paiement « **ouvert autant que possible, fermé autant que nécessaire...** »
- « **...rend obligatoire l'accès ouvert** pour les publications et pour les données issues de recherches sur projets »
- « **faire sortir la recherche** financée sur fonds publics du cadre confiné **des bases de données fermées**. Réduire les efforts dupliqués dans la collecte, la création, le transfert et la réutilisation du matériel scientifique. Augmenter ainsi l'efficacité de la recherche. »

Feuille de route CNRS pour la science ouverte (nov. 2019)

- « La **science ouverte** ne favorise pas seulement une approche transversale du partage des résultats de la science, ...
- ... En ouvrant les données, les processus, les codes, les méthodes ou encore les protocoles, elle offre **aussi une nouvelle façon de faire de la science.** »

1. Open Science : pourquoi FAIRe?

Plan données de la recherche du CNRS (nov 2020)

- « pour inciter les scientifiques à **rendre leurs données accessibles et réutilisables** »
- « le cadre légal des données de la recherche produites dans un contexte de financement sur fonds publics a évolué vers un principe de **mise à disposition en données ouvertes** »
- « incite toutes les structures de recherche à se doter de **politiques des données** ».

Les prospectives INSU (2021)

Défi 13 – de la production de données à leur exploitation scientifique

Défi 14 – accès ouvert aux données scientifiques

- Les données sont au cœur de la démarche scientifique de l'INSU
- Longue tradition de distribution ouverte des données via les SNOs et IR
- Une politique volontariste de soutien au partage de données:

**... Être le moteur de l'écosystème national de la donnée
en Sciences De l'Univers et Environnement**

2. Le cycle de la donnée

2. Le cycle de vie de la donnée

Le cycle de la donnée comprend l'ensemble des étapes par lesquelles passent les données scientifiques.

Étapes principales :

1. **Préparation** : montage du projet, planification
2. **Production** : capteurs, satellites, enquêtes, simulations
3. **Traitement** : nettoyage, transformation, calibration
4. **Documentation** : ajout de métadonnées normalisées
5. **Archivage et stockage** : dépôts institutionnels, bases de données
6. **Diffusion** : portails de données, API, plateformes
7. **Réutilisation** : analyses, visualisations, intelligence artificielle



Guide de bonnes pratiques sur la gestion des données de la recherche

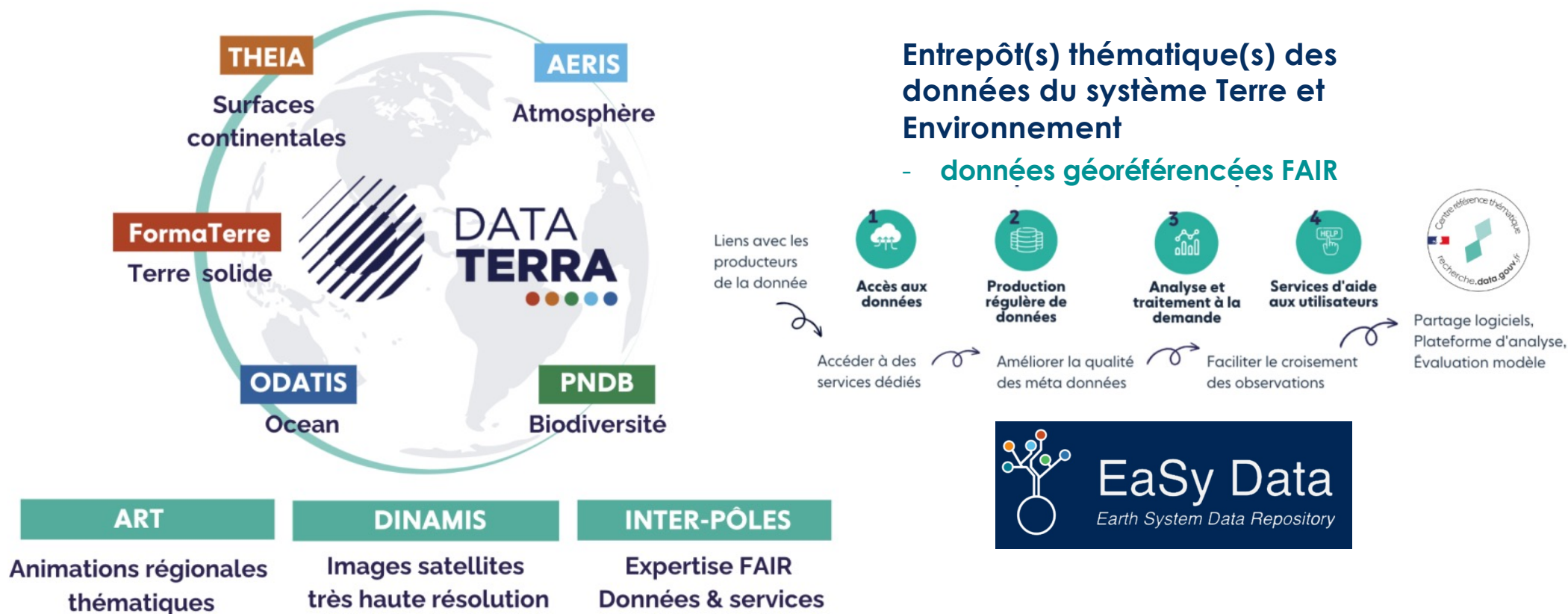
<https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html>



https://rdmkit.elixir-europe.org/data_life_cycle

2. Le cycle de vie de la donnée

DATA TERRA: E-Infrastructure de Recherche système Terre et Environnement



2. Le cycle de vie de la donnée

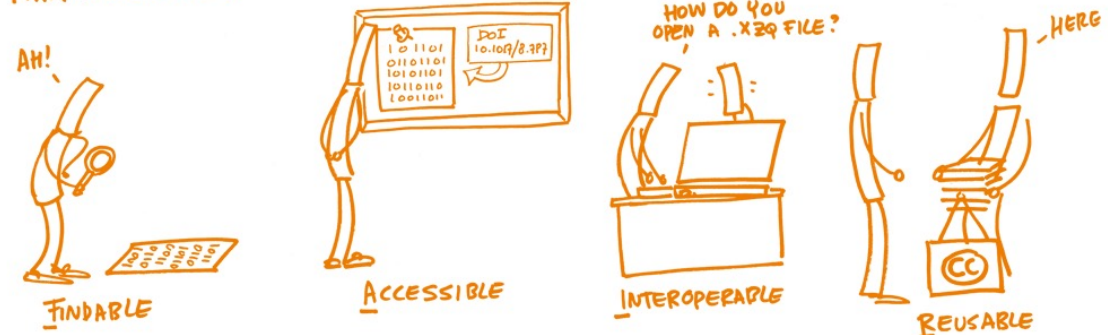
FAIRe : Findable, Accessible, Inter-operable, Re-usable, eco friendly

- **Facile à trouver** : cataloguer et référencer, un identifiant unique et persistant
- **Accessible** : services d'accès par l'humain ou la machine, extraction/agrégation
- **Interopérable** : machine à machine, vocabulaire et syntaxe communs, API standardisées,
- **Réutilisables** : répétable, reproductible, référencée, riche en description et méthodes
- **Ecoresponsable** : choix du média de stockage et localisation au plus proche



<https://www.force11.org/group/fairgroup/fairprinciples>

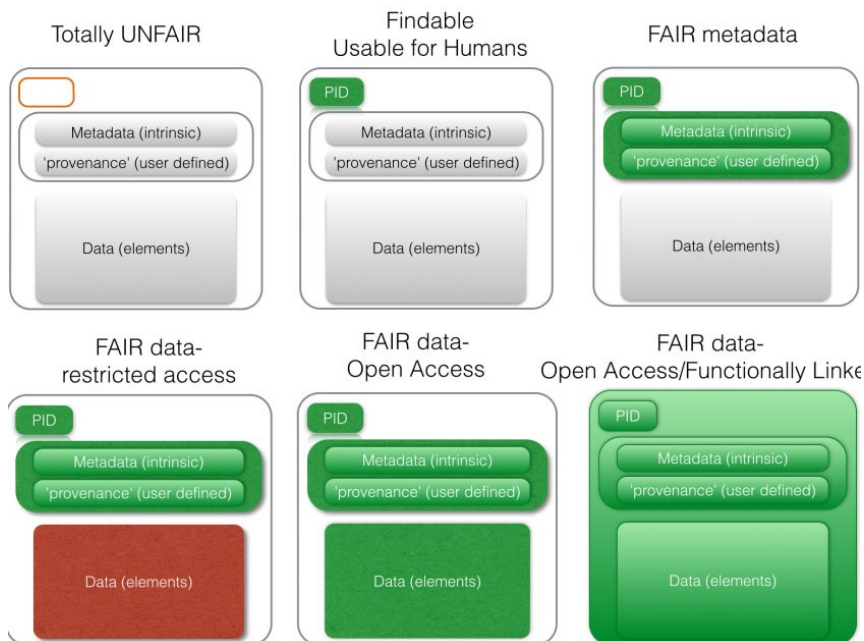
FAIR DATA PRINCIPLES



2. Le cycle de vie de la donnée

Le modèle de données distribuable

Data as increasingly FAIR Digital Objects



Niveau 1: Chaque objet de données a un **identifiant** machine **résoluble persistant et unique FAIR PID**

Niveau 2: Chaque objet de données est environné de **métadonnées** pour donner la provenance des données, ce qui lui est arrivé, qui l'a utilisé pour, et peut être utilisé pour, etc

➔ CATALOGAGE POSSIBLE

Niveau 3: Les données des éléments eux-mêmes dans les objets de données sont techniquement aussi FAIR, mais pas complètement libre accès et non réutilisables sans restrictions

Niveau 4: Les métadonnées ainsi que les éléments de données eux-mêmes sont entièrement FAIR et entièrement publique, **sous licence bien définie**

➔ OPEN ACCESS

2. Le cycle de vie de la donnée

Les formats de données distribuables

Autosuffisance du fichier de données :

la donnée distribuée contient tous les attributs pour la décrire, connaître le mode d'acquisition (par qui, quand, où, comment), et concernant la politique de diffusion (licence)

- Uniformiser le format de stockage des données issues des observations en géosciences
- Disposer d'un format auto-documenté (informations intrinsèques, pas de tables externes)

Formats de données spatiales multidimensionnelles :

- **GRIB** GRidded In Binary, Organisation Météorologique Mondiale
- **NetCDF** : Network Common Data Form, University Corporation for Atmospheric Research (UCAR) - **Climate and Forecast conventions (CF conventions)**
- **GeoTIFF** : Tag Image File Format
- **HDF** : Hierarchical data format, National Center for Supercomputing Application (NCSA) - **Earth Observing System conventions (EOS conventions)**

<https://data.ifremer.fr/Tout-savoir-sur-les-donnees/Gestion-des-donnees/Formats>

BONUS

Qu'est-ce que le NetCDF ?

NetCDF (Network Common Data Form) est constitué, d'une part, d'un ensemble de bibliothèques logicielles libres et gratuites et d'autre part, d'un format de données « auto-documenté » qui permet la création, l'accès et le partage de données scientifiques stockées sous la forme de tableaux.

Le fichier NetCDF (.nc) est un fichier binaire :

- **Auto-descriptif** : le fichier NetCDF file comprend des informations sur les données qu'il contient
- **Portable** : un fichier netCDF est accessible par n'importe quelle machine indépendamment de l'architecture matérielle (quelque soit le stockage des différents types entiers, flottants, caractères, ...)
- **Accès direct** : l'accès à une partie du fichier (sous-ensemble de tableau) est efficace
- **Agrégable** : Les données peuvent être agrégée dans un fichier NetCDF structuré sans recopier les données ou redéfinir la structure
- **Partagé** : un utilisateur peut accéder au fichier en écriture pendant que d'autres sont en train de le lire
- **Rétro-Compatibilité** : les formes antérieures de données netCDF sont compatibles avec les versions futures.

BONUS

Vocabulaire NetCDF de base

<https://desktop.arcgis.com/fr/arcmap/10.3/manage-data/netcdf/essential-netcdf-vocabulary.htm>

Dimensions : une dimension netCDF possède à la fois un nom et une taille.

Variables : une variable représente un ensemble de valeurs de même type.

Variables de coordonnées : une variable unidimensionnelle avec le même nom qu'une dimension

Attributs : les attributs netCDF (globaux ou variables) permettent de stocker des données auxiliaires ou des métadonnées.

Conventions : les conventions définissent des métadonnées qui fournissent une description définitive des données de chaque variable, ainsi que leurs propriétés spatiales et temporelles. Une convention aide les utilisateurs de données de sources différentes à déterminer les quantités qui sont comparables. Le nom de la convention est présenté en tant qu'attribut global dans un fichier netCDF.

BONUS

En tête d'un fichier NetCDF

```
ncdump -h data/HFRdata/HFR_MED_TLN_T0TL_20210925T0101Z.nc
```

dimensions:

```
TIME = 1 ;  
LATITUDE = 100 ;  
LONGITUDE = 108 ;  
STRING50 = 50 ;  
MAXINST = 1 ;  
STRING200 = 250 ;  
REFMAX = 1 ;  
DEPTH = 1 ;  
MAXSITE = 4 ;  
STRING15 = 15 ;
```

variables:

```
double TIME(TIME) ;  
TIME:long_name = "Time" ;  
TIME:standard_name = "time" ;  
TIME:units = "days since 1950-01-01T00:00:00Z" ;  
TIME:calendar = "Gregorian" ;  
TIME:axis = "T" ;  
TIME:sdn_parameter_name = "Elapsed time (since 1950-01-  
01T00:00:00Z)" ;  
TIME:sdn_parameter_urn = "SDN:P01::ELTJLD01" ;  
TIME:sdn_uom_name = "Days" ;  
TIME:sdn_uom_urn = "SDN:P06::UTAA" ;  
TIME:ancillary_variables = "TIME_QC" ;
```

```
double EWCT(TIME, DEPTH, LATITUDE, LONGITUDE) ;  
EWCT:_FillValue = -32767. ;  
EWCT:long_name = "West-east current component" ;  
EWCT:standard_name = "eastward_sea_water_velocity" ;  
EWCT:units = "m s-1" ;  
EWCT:scale_factor = 1. ;  
EWCT:add_offset = 0. ;  
EWCT:ioos_category = "Currents" ;  
EWCT:...
```

BONUS

En tête d'un fichier NetCDF

```
ncdump -h data/HFRdata/HFR_MED_TLN_TOTL_20210925T0101Z.nc
```

```
// global attributes:
```

```
:netcdf_version = "4.7.3" ;  
:netcdf_format = "netcdf4_classic" ;  
:site_code = "HFR-MedTln" ;  
:platform_code = "HFR-MedTln-Total" ;  
:data_mode = "R" ;  
:DoA_estimation_method = "POB: Direction Finding; PEY: Direction Finding" ;  
:calibration_type = "POB: internal; PEY: internal" ;  
:last_calibration_date = "2021-09-25T09:51:33Z" ;  
:calibration_link = "Dylan Dumas, Charles-Antoine Guerin, Self-calibration and  
antenna grouping for bistatic oceanographic High-Frequency Radars,2020,  
https://arxiv.org/abs/2005.10528" ;  
:title = "HFR-MIO-CURRENTS" ;  
:summary = "Sea surface currents recorded by High Frequency (HF) radar in the  
north-western Mediterranean Sea. The network has been in situ on the French  
Riviera since 2012 and is operated by MIO - the Mediterranean Institute of  
Oceanography (UMR Aix Marseille Université, CNRS, IRD, Université de Toulon)."  
:keywords = "Earth Science > Oceans > Ocean Circulation > Ocean currents, Earth  
Science > Oceans > Ocean Circulation > Wind-driven circulation, In Situ Land-  
based Platforms > Observatories, In Situ Land-based Platforms > Radio  
transmitters, ocean > sea_surface_skin, OCEAN > ATLANTIC OCEAN > NORTH ATLANTIC  
OCEAN > MEDITERRANEAN SEA, Earth Remote Sensing Instruments > Active Remote  
Sensing > Imaging Radars" ;  
:keywords_vocabulary = "GCMD Science Keywords" ;
```

3. Métadonnées et vocabulaire contrôlé

□ Métadonnée

Une **métadonnée** est une donnée qui décrit une autre donnée.
Elle fournit des informations contextuelles, techniques ou sémantiques pour comprendre, retrouver, utiliser ou préserver la donnée décrite.

Exemple :

Pour un fichier de mesure de **température**, les métadonnées peuvent inclure :

- **où** : la localisation géographique,
- **quand** : la date,
- **comment** : le capteur utilisé,
- **quoi** : la température de l'eau
- l'unité,
- l'incertitude de mesure
- la plage de mesures [min; max]

3. Métadonnée & vocabulaire contrôlé

□ Vocabulaire contrôlé

Un **vocabulaire contrôlé** est un ensemble normalisé et validé de termes ou concepts, utilisés pour décrire les données de manière cohérente et non ambiguë.

Il sert à :

Harmoniser la terminologie entre disciplines ou institutions,
Faciliter la recherche et l'interopérabilité,
Réduire l'ambiguïté (ex. "CO₂" vs "dioxyde de carbone").

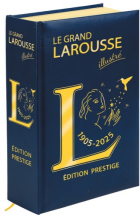
On peut distinguer plusieurs niveaux :

- **Listes de termes / codes** (p. ex. codes de variables du climat),
- **Thésaurus** (relations hiérarchiques et associatives entre termes),
- **Ontologies** (modèles plus riches avec relations logiques et formelles, exploitables par les machines).



3. Métadonnée & vocabulaire contrôlé

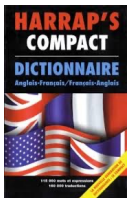
□ Quelques références



Thésaurus

GEMET (General Multilingual Environmental Thesaurus) : thésaurus multilingue de l'Agence européenne pour l'environnement (EEA). <https://www.eionet.europa.eu/gemet>

INSPIRE thesauri : vocabulaires recommandés pour la directive européenne INSPIRE (informations environnementales géographiques)



Ontologies

Earth Portal Le catalogue des ontologies et des artefacts sémantiques du système Terre, de l'environnement et des domaines connexes <https://earthportal.eu/ontologies>

 **Standards de métadonnées environnementales** (associés à vocabulaires contrôlés) :
ISO 19115 / 19139 (géoinformation – métadonnées), largement utilisé pour les données spatiales.
Dublin Core (générique, interdisciplinaire).
INSPIRE (2007) Metadata Implementing Rules (Europe).
CF Conventions (Climate and Forecast) pour les fichiers NetCDF en climatologie.

3. Métadonnée & vocabulaire contrôlé



Natural
Environment
Research Council



National Oceanography Centre
British Oceanographic Data
Centre BODC

The NERC Vocabulary Server (NVS)

Service Status

[NVS Home](#) | [Vocabularies](#) | [Thesauri](#) | [Search NVS](#) | [SPARQL](#) | [Other Tools](#) | [About NVS](#)

Concept

Alternate Formats

Partager un vocabulaire et une logique commune

Selon le **World Wide Web Consortium W3C**, « le Web sémantique fournit un modèle qui permet aux données d'être partagées et réutilisées entre plusieurs applications, entreprises et groupes d'utilisateurs »

Le modèle sémantique est basé sur la conceptualisation de ce qui constitue une mesure et l'atomisation en ses différents constituants

la propriété d'un objet dans un contexte définie par une méthode

<https://earthportal.eu/>

3. Métadonnée & vocabulaire contrôlé

la propriété d'un objet dans un contexte définie par une méthode

exemple :
lexique et
vocabulaire

Speed of water current (Eulerian measurement) in the water body by directional range-gated radar

RADIAL SEAWATER VELOCITY	
URI	URI
Within Vocab	Within Vocab
Alternative Labels	Alternative Labels
Definition	Definition
Date	Date
Identifier	Identifier
Note	Note
Has Current Version	Has Current Version
Version	version
version	version
Same As	Same As
Broader	Broader
	Related

Radial sea water velocity away from instrument

URI	http://vocab.nerc.ac.uk/standard_name/radial_sea_water_velocity_away_from_instrument/	
Within Vocab	Climate and Forecast Standard Names	
Definition	A velocity is a vector quantity. "Radial velocity away from instrument" means the component of the velocity along the line of sight of the instrument where positive implies movement away from the instrument (i.e. outward). The "instrument" (examples are radar and lidar) is the device used to make an observation. A standard name referring to radial velocity "toward_instrument" should be used for a data variable having the opposite sign convention.	
Date	None	
Note	accepted	
modified	2019-06-17 13:47:18.0	
Broad Match	P02:RFVL	Horizontal velocity of the water column (currents)
Exact Match	http://vocab.nerc.ac.uk/standard_name/radial_sea_water_velocity_away_from_instrument/ P07:CFV13N15 radial_sea_water_velocity_away_from_instrument http://mmisw.org/ont/cf/parameter/radial_sea_water_velocity_away_from_instrument	
Related	P06:UVAA	Metres per second

3. Métadonnée & vocabulaire contrôlé

❑ Convention « Climate Forecast »

<http://cfconventions.org/> défini les standards pour la communauté océan et atmosphère
 minimiser les erreurs « stupides » ex : Celsius ou Fahrenheit, pied ou mètre, nœud ou mètre/seconde

Origine de la donnée

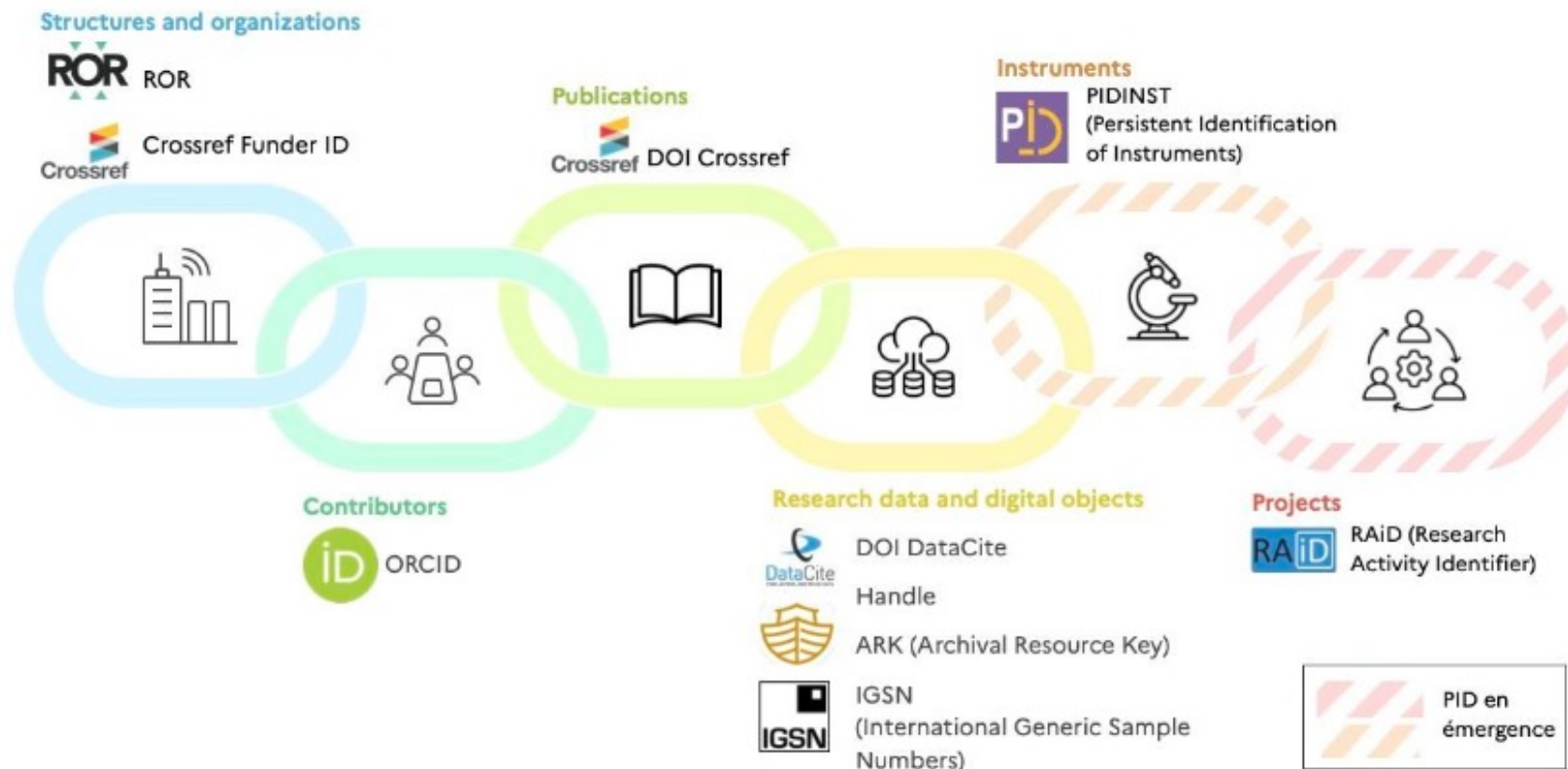
title	Qui-a-t-il dans le fichier ?
institution (*)	Où a-t-il été produit ?
source (*)	Comment a-t-il été produit ? (version du modèle, de l'instrument, etc ...)
history	Protocole de traitement
references (*)	Publication de référence, documentation, ...
comment (*)	Commentaires divers

Description de la donnée

units	Unité internationale (degC, Pa, m.s ⁻¹ , ...)
standard_name	Identifier la donnée ex : sea_surface_temperature
long_name	Pas de standard ex : Sea Surface Temperature
_FillValue et/ou missing_value	Valeur des données manquantes ex : 999.9
valid_min et valid_max	le min et le max acceptable de la variable
add_offset et scale_factor	facteur d'échelle et valeur de référence ex : SST = 0.f + 1.0f * value
flag_values et/ou flag_meaning	si la donnée nécessite une valeur de statut ex: quality_good , outside_valid_range

3. Métadonnée & vocabulaire contrôlé

□ un identifiant unique et pérenne : PID



4. OGC SensorThings



Open Geospatial Consortium.

□ le standard *sensorThings*

Standard de l'OGC **SensorThings** [[*OGC SensorThings API Part 1 - Sensing*, 2018](#)]

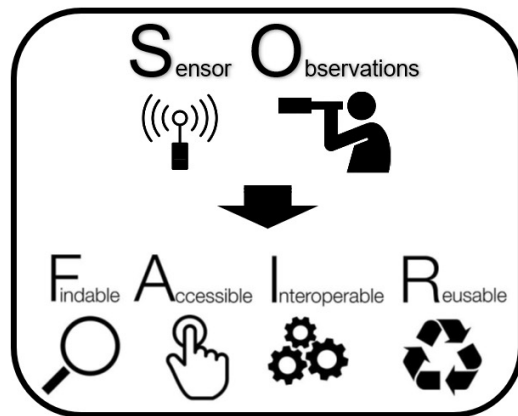
permet de publier les différentes composantes d'un observatoire (points de mesure, capteurs, paramètres observés, mesures, géolocalisation, chroniques...) conformément au principe du FAIR data. Il repose sur REST et JSON et est interopérable avec d'autres standards OGC

- **Thing** : objet mesuré (ex. station météo)
- **Sensor** : dispositif de mesure
- **ObservedProperty** : phénomène mesuré (ex. température, humidité)
- **Datastream** : flux de données
- **Observation** : valeur mesurée à un instant donné

4. OGC SensorThings

□ sensorThings pour les nuls ...

<https://geosass.fr/sofair-book/intro.html>

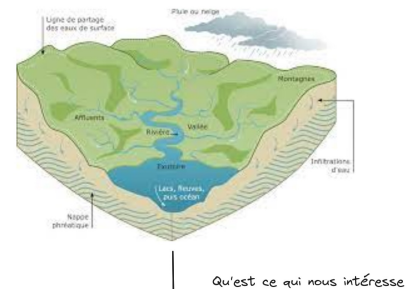


Le SOFAIR book

par Mario Adam¹, Stéphane Bujacic¹, Manuel Chev², Paloma Cucchi¹, Mikaël Fauchoux¹, Ophélie Fovet¹, Christophe Geneste¹, Tom Loree¹, Olivier Maury⁴, Thierry Morvan¹, Pascal Pichelin¹, Paul Robin¹, François Rouault¹, Laurent Ruiz³, Hervé Squidant¹, Zahra Thomas¹

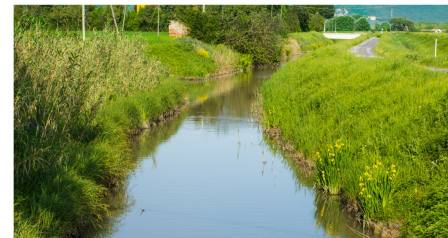
Qu'est ce qu'on va étudier

Un bassin versant



Qu'est ce qui nous intéresse ?

L'exutoire du BV (Thing)



Que veut on observer ?
un phénomène physique
des variables chimiques ?

Le débit (Observed property)



Comment l'observe t'on ?
Quel instrument ?

Station hydro (Sensor)



Production de série temporelle (Datastream)

Débit (Observation)



□ *SensorThings API*

Une **API** (application programming interface ou « interface de programmation d'application ») est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

Les spécificités de l'API SensorThings sont :

- une **API RESTful** pour les interactions entre les services et le client,
- le formalisme des requête conforme au standard Open Data Protocol (**OData**),
- l'encodage des données au format JavaScript Object Notation (**JSON**),
- une extension du protocole de messagerie pour la publication et l'abonnement disponible pour les opérations en temps réel, basée sur la norme ISO Message Queuing Telemetry Transport (**MQTT**),
- un système de **pagination** du résultat des requêtes,
- la capacité de **découverte** à l'aide d'un simple navigateur web.

5. Démonstration

Service de téléchargement des données Site Urbain Rennais (SUR), projet City Orchestra



Cette API diffuse des données hydrologique et climatique, à partir de d'octobre 2021, au standard OGC SensorThings. La zone d'étude est le quartier de Beauregard au nord-ouest de Rennes

❑ *Le catalogue*

<https://geodata.inrae.fr/datahub/service/ff3ae11e-a774-4ca0-9eac-e966f3df82d5>

<https://geodata.inrae.fr/geonetwork/srv/fre/catalog.search#/metadata/ff3ae11e-a774-4ca0-9eac-e966f3df82d5>

Service de téléchargement des données Site Urbain Rennais (SUR), projet City Orchestra



Cette API diffuse des données hydrologique et climatique, à partir de d'octobre 2021, au standard OGC SensorThings. La zone d'étude est le quartier de Beauregard au nord-ouest de Rennes

❑ Découverte du service

<https://geosas.fr/cityorchestra/vidae-2-accueil.html>

6. Conclusions

Open Science : comment FAIRe?



- Partager les informations sur les outils et les ressources de mise en œuvre
- Contextualiser la mise en œuvre
- La granularité de l'information fait qu'elle est compréhensible
- Même hors de son contexte
- L'information est enregistrée dans un entrepôt de données
- Et il existe une description des informations à travers des métadonnées spécifiques
- Qui sont interrogeables à travers un catalogue
- Sont interoperables dans le sens où l'information peut être échangée
- Est réutilisable par d'autres
- Peut être réutilisée pour différents contextes
- Peut être assemblée

Skills4EOSC 'Skills for the European Open Science commons: creating a training ecosystem for Open and FAIR science'

FAIRe de l'Open Science



Guide de bonnes pratiques
sur la gestion des données
de la recherche

<https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html>



Plan de gestion de la donnée



Infrastructure d'hébergement de la donnée



Donnée documentée (metadata)



Conventions, nomenclatures



Outils existants

Merci de votre attention