Diagnostic sur la gestion et la mise à disposition des données des observatoires

Auteurs : Isabelle Braud, Sylvie Galle, Véronique Chaffard, Charly Cousot (Theia/OZCAR)

Version : 2021-06-11

Contexte et objectif de ce document

Les services d'observation labellisés doivent mettre les données collectées à disposition de la communauté scientifique. C'est pourquoi les observatoires ont développé des sites web, des portails de données ou tout autre moyen de donner accès aux données qu'ils produisent.

Par ailleurs, la gestion des données est en train de se professionnaliser avec la promotion des principes des données FAIR (Findable, Accessible, Interoperable, Reusable, Wilkinson et al., 2016¹) qui ont été formalisés par l'initiative GO-FAIR² et la mise en place de Plan de Gestion des Données (PGD). Le CIRAD³ ou INRAE⁴ proposent une page web bien documentée pour aider à réfléchir à la gestion et au partage des données. Voir définitions des termes en annexe 1.

Quels critères pour savoir si des données sont FAIR ?

Un groupe de travail du RDA (Research Data Alliance), le « Fair data maturity model working group »⁵ a travaillé sur une harmonisation d'une méthodologie pour définir si des données respectent les principes FAIR. Un travail a permis de comparer différentes méthodologies⁶. Ce document renvoie à plusieurs tentatives de définition de critères pour juger du degré de respect des principes FAIR. Il a débouché sur un document de recommandation et une liste de critères et de manière de les évaluer qui peut être consulté ici⁷. Les critères retenus sont donnés en annexe 2.

Qu'en est-il des données de votre observatoire ?

Les critères proposés pour évaluer si des données sont FAIR peuvent paraître un peu abstraits et leur implémentation ne relève pas nécessairement du niveau d'un observatoire pour tous ces critères.

C'est pourquoi Theia/OZCAR propose un **formulaire d'autodiagnostic** des pratiques de gestion des données des observatoires. Ce formulaire s'appuie sur des questions simples et concrètes qui devraient être plus parlantes pour des personnes pas forcément spécialistes de la gestion des données. Il est formulé dans une feuille Excel (**FAIR_test_V5.xlsx**).

¹ Wilkinson, M.D., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18

² <https://www.go-fair.org/fair-principles/>

³ <https://coop-ist.cirad.fr/gerer-des-donnees/s-initier-en-ligne-aux-donnees-de-la-recherche/1-familiarisez-vous-avec-le-concept-de-donnees-de-la-recherche>

⁴ <https://datapartage.inrae.fr/Produire-des-donnees-FAIR>

⁵ <https://www.rd-alliance.org/groups/fair-data-maturity-model-wg>

⁶ <https://www.rd-alliance.org/system/files/Results%20Analysis%20of%20FAIR%20assessment%20tools%20v3.pdf>

⁷ <https://zenodo.org/record/3909563#.YKE9FOfgpPZ>

Objectif du formulaire d'autodiagnostic :

Ces questions sont là pour vous **aider à identifier vos points forts et les actions qu'il serait nécessaire de faire pour aller vers des données les plus FAIR possibles**. Certains items reprennent en plus des questions posées dans les plans de gestion des données.

Par ailleurs, il existe maintenant un niveau de gestion des données via le Système d'Information Theia/OZCAR, qui propose différents services standardisés pour la découverte et l'accès aux données de vos observatoires. Le SI Theia/OZCAR peut donc accroître la visibilité de vos données, notamment sur l'interopérabilité. Le questionnaire proposé permet **d'identifier tous les services que Theia/OZCAR offre ou offrira⁸ et que vous pouvez lui déléguer**.

Comment utiliser le formulaire

Le document Excel propose un diagnostic sur la manière dont les données des observatoires sont actuellement gérées et mises à disposition. Il s'agit d'une grille d'analyse des pratiques actuelles en termes de documentation des observatoires, des données qu'ils produisent et de leur mise à disposition.

Vous reconnaîtrez dans les items du questionnaire une partie des informations (métadonnées) demandées par le format pivot de Theia/OZCAR. Les métadonnées obligatoires, mais aussi les métadonnées « recommandées » : elles sont toutes utiles pour une meilleure réutilisabilité de vos données. La colonne A du questionnaire indique quelles sont les fonctionnalités et critères FAIR qui sont remplis dès lors que votre observatoire a mis en place le modèle pivot Theia/OZCAR et est visible sur le portail de données.

Les différentes questions ont été rattachées à une des lettres du FAIR ou à une catégorie « D » plus organisationnelle sur la gestion et la mise à disposition des données.

Il y a quatre séries de questions concernant :

1. Des questions générales sur la gestion des données de l'observatoire (aspects organisationnels)
2. Des questions sur le producteur des données (l'observatoire)
3. Des questions sur les observations (chroniques, cartes, etc..)
4. Des questions sur les jeux de données disponibles sur le web (si l'observatoire en dispose)

Pour chaque question, les réponses attendues sont oui/non/partiellement. En fonction de la réponse, des précisions peuvent être demandées.

A partir des réponses, une notation est proposée qui permet de positionner vos pratiques actuelles par rapport aux critères FAIR et D proposés dans le questionnaire.

Remarques à propos du score :

- 1- Certains items amènent une question supplémentaire. La réponse à cette question est obligatoire pour obtenir les points.
- 2- Certains critères du FAIR (mentionnés en Annexe 2) ne sont pas pris en compte par ce questionnaire car ils relèvent plutôt du SI Theia/OZCAR. Les notes obtenues sur les différents critères FAIR/organisation sont donc incomplètes.

Le diagnostic permet cependant à chaque observatoire de mesurer la trajectoire qu'il doit emprunter pour progresser vers des données plus FAIR, et la valeur ajoutée qu'apportera la mise en visibilité de ses données sur le SI Theia/OZCAR (colonne A du fichier Excel et annexe 4 de ce document).

⁸. Certains nouveaux services sont en cours de développement.

Annexes

Annexe 1 : Quelques définitions

Annexe 2 : Les indicateurs du FAIR selon le Fair data maturity model du RDA

Annexe 3 : Quelques éléments sur le choix de licence sur les jeux de données

Annexe 4 : Dossier à remplir pour le bilan et la labellisation d'un SNO

Références

Annexe 1 : Quelques définitions

- **Donnée** : c'est une information, une observation, une mesure,
- **Métadonnées** : elles fournissent des informations sur le contexte de la donnée (des données sur la donnée). Une donnée sans métadonnées est inexploitable en dehors de ceux qui l'ont acquise. Les métadonnées peuvent être plus ou moins riches (complètes).
- **Jeux de données (data set)** : un jeu de données est l'agrégation d'enregistrements de données organisés pour former un ensemble cohérent. Un jeu de données peut se concevoir de plusieurs points de vue : question scientifique, objet d'intérêt, responsable(s) du jeu de données, traitements, échelles spatiales et temporelles explorées, des types de capteurs (réseau d'instruments).
C'est en général au niveau d'un jeu de données qu'on attribue un DOI (Digital Object Identifier), ce qui nécessite de bien réfléchir à la granularité de ces jeux de données, puisque chaque jeu de données doit être accompagné des métadonnées adaptées.
- **Open Data**⁹
Les données ouvertes ou Open Data sont des données numériques dont l'accès et l'usage sont laissés libres aux usagers (public, privé, citoyen), produites notamment par une collectivité ou un établissement public. Elles sont diffusées de manière structurée selon une méthode et une licence ouverte garantissant leur libre accès et leur réutilisation par tous, sans restriction technique, juridique ou financière.
Les données en Open data peuvent donc être utilisées, réutilisées et redistribuées librement.
Mais l'Open Data, ce sont des données « aussi ouvertes que possibles, aussi fermées que nécessaires », ce qui autorise autant d'exceptions à l'ouverture (protection des intérêts de l'Etat, des droits de la personne humaine et des biens tels que : sécurité publique, sûreté nationale, confidentialité, secrets d'affaires, droits de propriété intellectuelle de tiers).
- **Attention : FAIR data n'est pas synonyme de Open Data**¹⁰.
Des données peuvent être FAIR sans pour autant être disponibles librement. Ce qui importe dans le FAIR c'est d'apposer une licence bien définie aux données. Ces licences peuvent restreindre l'usage des données (si elles incluent des données personnelles par exemple « aussi fermé que nécessaire »), l'essentiel étant que la licence attribuée aux données soit claire, et de préférence qu'elle puisse être reconnue par des machines.
- **Plan de Gestion des Données**¹¹ :
Un plan de gestion des données (PGD) est un document formalisé, établi au démarrage d'un projet de recherche, et qui décrit la façon dont les données produites seront obtenues, documentées, utilisées, gérées, sauvegardées et disséminées, pendant la phase de recherche et une fois le projet terminé.
Le plan de gestion des données prend en compte toutes les étapes du cycle de vie des données et doit être mis à jour tout au long du projet de recherche.
L'ANR propose un template pour rédiger un Plan de Gestion des Données.

⁹ https://fr.wikipedia.org/wiki/Donn%C3%A9es_ouvertes

¹⁰ <https://www.go-fair.org/resources/faq/ask-question-difference-fair-data-open-data/>

¹¹ Définition tirée de: <https://datapartage.inrae.fr/Gerer/Rediger-un-PGD>

- **Données FAIR**

Les principes du FAIR s'appliquent à trois entités: les données (ou tout objet numérique), les métadonnées (les informations sur l'objet digital), et l'infrastructure où les données sont déposées.

- **F-Findable**¹² : les données doivent être faciles à trouver, par les humains et les machines. Au-delà de pouvoir être trouvable sur le web, le F du FAIR insiste sur l'existence d'identifiants uniques qui permettent de rendre l'information plus facile à trouver, en particulier par les machines (ces identifiants sont l'équivalent des DOIs sur les articles ou les jeux de données)
F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata
F3. Metadata clearly and explicitly include the identifier of the data they describe (par exemple, on trouve le numéro du DOI du jeu de données décrites par les métadonnées dans les métadonnées)
F4. (Meta)data are registered or indexed in a searchable resource
- **A-Accessible** : les données doivent pouvoir être téléchargeables
A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorization procedure, where necessary
A2. Metadata are accessible, even when the data are no longer available
- **I- Interoperable** : Le terme interopérabilité¹³ désigne le fait que des systèmes informatiques puissent interagir entre eux, afin de créer un réseau et de faciliter le transfert de données. On distingue l'interopérabilité technique (pouvoir communiquer), sémantique (pouvoir se comprendre) et syntaxique (savoir communiquer). Elle implique l'utilisation de protocoles d'échanges standards, des métadonnées contextuelles précises, et de contenu et de formats respectant des standards internationaux et utilisant un vocabulaire respectant lui-même les principes FAIR
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data
- **R- Reusable**: les données doivent être réutilisables par d'autres utilisateurs. Ceci doit être facilité par l'utilisation de standards communs, grâce à des bases de données rassemblant des données claires, vérifiées et bien décrites. Les (méta)données doivent être richement décrites, par une pluralité d'attributs précis et pertinents, incluant des détails sur leur provenance (c'est-à-dire sur l'historique des données et les protocoles et traitements qui ont permis de l'obtenir). Elles sont fournies avec une licence d'utilisation claire et accessible (c'est-à-dire trouvable sur le web) (voir annexe 2 pour les recommandations Theia/OZCAR).
R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (Meta)data are released with a clear and accessible data usage license
R1.2. (Meta)data are associated with detailed provenance
R1.3. (Meta)data meet domain-relevant community standards

¹² Pour des explications en français, voir les définitions proposées par Wikipedia : https://fr.wikipedia.org/wiki/Fair_data et la description détaillée de chacun des principes sur le site Go-FAIR : <https://www.go-fair.org/fair-principles/>

¹³ https://fr.wikipedia.org/wiki/Interop%C3%A9rabilit%C3%A9_en_informatique

Annexe 2 : Les indicateurs du FAIR selon le Fair data maturity model du RDA

Ce tableau reprend les 41 indicateurs proposés par le working group du RDA¹⁴. Ils sont classés en essentiels, importants et utiles.

FAIR	ID	Indicator	Priority
F1	RDA-F1-01M	Metadata is identified by a persistent identifier	●●● Essential
F1	RDA-F1-01D	Data is identified by a persistent identifier	●●● Essential
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	●●● Essential
F1	RDA-F1-02D	Data is identified by a globally unique identifier	●●● Essential
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	●●● Essential
F3	RDA-F3-01M	Metadata includes the identifier for the data	●●● Essential
F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	●●● Essential
A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	●● Important
A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	●●● Essential
A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	●●● Essential
A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	●●● Essential
A1	RDA-A1-03D	Data identifier resolves to a digital object	●●● Essential
A1	RDA-A1-04M	Metadata is accessed through standardised protocol	●●● Essential
A1	RDA-A1-04D	Data is accessible through standardised protocol	●●● Essential
A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	●● Important
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	●●● Essential
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	●● Important
A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	● Useful
A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	●●● Essential
I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	●● Important
I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	●● Important
I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	●● Important
I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	●● Important
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	●● Important
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	● Useful
I3	RDA-I3-01M	Metadata includes references to other metadata	●● Important
I3	RDA-I3-01D	Data includes references to other data	● Useful
I3	RDA-I3-02M	Metadata includes references to other data	● Useful

¹⁴ <https://zenodo.org/record/3909563#.YKE9FOfgpPZ>

FAIR	ID	Indicator	Priority
I3	RDA-I3-02D	Data includes qualified references to other data	● Useful
I3	RDA-I3-03M	Metadata includes qualified references to other metadata	●● Important
I3	RDA-I3-04M	Metadata include qualified references to other data	● Useful
R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	●●● Essential
R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	●●● Essential
R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	●● Important
R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	●● Important
R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	●● Important
R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	● Useful
R1.3	RDA-R1.3-01M	Metadata complies with a community standard	●●● Essential
R1.3	RDA-R1.3-01D	Data complies with a community standard	●●● Essential
R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	●●● Essential
R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	●● Important

Annexe 3 : Quelques éléments sur le choix de licence sur les jeux de données

Avant la publication de données, il est nécessaire d'apposer une licence sur le jeu de données (recommandations des juristes) qui précise comment le jeu de données peut être réutilisé et quels sont les devoirs des personnes qui les réutilisent. Ces licences ont une valeur juridique, contrairement à une charte d'utilisation de données. Les licences protègent les fournisseurs de données contre une mauvaise utilisation de leurs données.

Concernant les données des SNOs, il est attendu que les données soient diffusées librement à tout utilisateur et une licence de type Creative Commons Attribution¹⁵ est recommandée. Il existe plusieurs types de licences Creative Commons. Nous recommandons l'usage de la licence CC-BY 4.0 qui permet un usage libre des données moyennant la citation des sources.

Pour faire reconnaître votre travail sur la production des données, il est préférable de demander aux utilisateurs de citer le DOI sur le jeu de données, ce qui permet un traçage plus aisé de l'utilisation et de la citation de vos données.

Annexe 4 : Dossier de labellisation à remplir pour l'INSU (2021)

Proposition de réponse pour les SNO d'OZCAR

A15- Lien avec un pôle ou un centre de données (oui/non) : oui

Si oui, préciser le(s)quel(s) (inclure une preuve d'appartenance ou d'adhésion en annexe au dossier) :

Depuis 2017, le pôle de données Theia et l'IR OZCAR développent ensemble le système d'information (SI) des surfaces continentales Theia/OZCAR (Braud et al., 2020) qui fait partie de l'IR Data Terra. Ce SI est opérationnel depuis le 01/01/2020 (<https://in-situ.theia-land.fr/>). Ce SI ne se substitue pas aux portails de données existants mais il les met en visibilité avec une sémantique unique. En 2021 les jeux de données d'un tiers des observatoires d'OZCAR sont visibles en temps réel sur le portail

¹⁵ <https://creativecommons.org/licenses/?lang=fr-FR>

Theia/OZCAR. A terme tous les observatoires seront visibles, mais ils doivent pour cela écrire un script qui extrait les données et les informations qui les décrivent dans le format pivot Theia/OZCAR.

Ajouter une précision sur votre SNO suivant son cas :

Le SNO XX est visible sur Theia/OZCAR depuis le JJ/MM/YYYY.

Le SNO XX planifie d'être visible sur le portail Theia/OZCAR d'ici xx an.

Le SNO XX n'est pas visible sur le portail Theia/OZCAR car nous n'avons pas les RH nécessaires.

Références :

Braud, I., Chaffard, V., Cousot, C., Galle, S., Juen, P., Alexandre, H., Baillion, P., Battais, A., Boudevillain, B., Branger, F., Brissebrat, G., Cailletaud, R., Cochonneau, G., Decoupes, R., Desconnets, J.-C., Dubreuil, A., Fabre, J., Gabillard, S., Gérard, M.-F., Grellet, S., Herrmann, A., Laarman, O., Lajeunesse, E., Le Hénaff, G., Lobry, O., Mauclerc, A., Paroissien, J.-B., Pierret, M.-C., Silvera, N., and Squidant, H.: Building the information system of the French Critical Zone Observatories network: Theia/OZCAR-IS, Hydrological Sciences Journal, 1–19, <https://doi.org/10.1080/02626667.2020.1764568>, 2020.

Results of an Analysis of Existing FAIR assessment tools (RDA) : <https://www.rd-alliance.org/system/files/Results%20Analysis%20of%20FAIR%20assessment%20tools%20v3.pdf>

Liens sur certains des outils ou publications examinées par le groupe RDA sur l'évaluation du FAIR

Checklist to evaluate FAIRness of data(sets (Note sur 13) :

<https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6ID5KuAz7yn0I0Opw/viewform>

- A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets : <https://datascience.codata.org/articles/abstract/10.2481/dsj.14-049/>
- FAIR Cook-Book: <https://fairplus.github.io/the-fair-cookbook/content/home.html>
- The WMO Stewardship Maturity Matrix for Climate Data (SMM-CD) : https://figshare.com/articles/dataset/The_WMO-Wide_Stewardship_Maturity_Matrix_for_Climate_Data/7006028
- Comment opérationnaliser et évaluer la prise en compte du concept 'FAIR' <https://zenodo.org/record/1995646#.XA-wpXRKg2w%C2%A0>
- FAIR Data Maturity Model. Specification and Guidelines <https://zenodo.org/record/3909563#.YKE9FOfgpPZ>