

FAIR: guide des bonnes pratiques

(prototype. version 3.1: 03/01/2022)

2022

Consortium COSME²

Site web: <https://cosme-2.github.io>

License: CC BY-NC-ND 4.0



Table des matières

1 Présentation des principes FAIR	3
1.1 Qu'est-ce que FAIR?	3
1.2 Ce que FAIR n'est pas	3
1.3 Pourquoi FAIRiser les données?	4
1.4 Qui FAIRise quoi?	6
1.5 Des données au dépôt de données: quelques notions principales ..	8
2 Étapes de la FAIRisation	18
Bibliographie	22
Annexe 1. Principes FAIR	24
Annexe 2. Principes FAIR étendus et les différents acteurs.....	26
Annexe 3. Choix d'un dépôt de données	29
Annexe 4. Modèle (provisionnel) des métadonnées.....	30
Annexe 5. Évaluation complète (questionnaire détaillé) de la correspondance des données aux principes FAIR	34
Annexe 6. Évaluation rapide (check-list) de la correspondance des données aux principes FAIR	37
Annexe 7. FAIR par la pratique & Éléments-clés des principes FAIR	38
Annexe 8. Exemple de FAIRisation par l'intermédiaire du versement de données dans un dépôt	45

1 Présentation des principes FAIR

1.1 Qu'est-ce que FAIR?

FAIR est une abréviation de "**Findable (= Facile à trouver), Accessible, Interoperable (= Interopérable), Reusable (= Réutilisable)**". Il s'agit de principes directeurs ou de recommandations (et non pas d'un standard) qui doivent être appliqués aux données de la recherche. Le but de la mise en place des principes FAIR est de rendre les données scientifiques plus accessibles et plus compréhensibles et leurs sauvegarde et réutilisation plus pérennes.

Les principes FAIR ont une visée multidisciplinaire; ils doivent être applicables à toutes les données de la recherche, de la physique nucléaire aux études littéraires. Cela a imposé certaines limites dans la formulation de ces principes. De fait, pour pouvoir être utilisables dans des domaines très divers, ces principes doivent rester les plus généraux possible. Ainsi, l'interprétation de ces principes et surtout leur application concrète restent le plus souvent à l'appréciation du responsable de chaque projet et peuvent donc varier, de façon considérable, d'un projet à l'autre, d'une discipline à l'autre.

Un autre écueil dans la mise en oeuvre des principes FAIR est leur relative nouveauté; FAIR et ses principes n'ont été définis qu'en 2016 [18]. De ce fait, le travail sur l'amélioration et le développement de ces principes, ainsi que sur de différents outils de FAIRisation, se poursuit toujours. À l'heure actuelle, il peut parfois s'avérer délicat de trouver des exemples concrets de données qui suivent fidèlement tous les principes FAIR (quel que soit le domaine de la recherche auquel ces données appartiennent).

1.2 Ce que FAIR n'est pas

FAIR data ne signifie pas Open Data ou Free Data. Les données peuvent être protégées contre la réutilisation et même contre un simple accès (s'il s'agit, par exemple, de données sensibles, classifiées ou privées). Un embargo (accessible dans X années) peut également être imposé aux données. Dans le cadre de la mise en oeuvre des principes FAIR, les données doivent être "*ouvertes autant que possible et fermées autant que nécessaire*" [11].

FAIR data ne signifie pas “archivage à vie”. La sauvegarde et l'archivage des données sont étroitement liés au lieu où les données sont déposées (dépôts institutionnels ou publics, centres de données privés, serveurs locaux, etc.). Cela peut induire des coûts supplémentaires et le financement de l'archivage des données à très long terme n'est jamais certain. En outre, les dépôts ou les centres de données eux-mêmes peuvent disparaître avec le temps. Ces différentes raisons peuvent conduire à la disparition des données. Ce problème peut être pallié, à certains égards, par l'indexation uniquement des métadonnées (cf. *infra* Annexe 1, principe FAIR:A2. Ce principe ne précise pas toutefois où ces seules métadonnées doivent être stockées).

1.3 Pourquoi FAIRiser les données?

Le travail avec les données numériques n'est pas toujours aisé et plusieurs obstacles et problématiques peuvent apparaître. Chacun des principes FAIR vise à répondre à un de ces problématiques.

Compréhension des données. Il n'est pas rare qu'une fois que l'article ou le projet est terminé, les données utilisées ou produites se retrouvent encodées dans un format dont on a déjà oublié le nom qui ne s'ouvre qu'avec un logiciel spécifique qui ne fonctionne plus que sur des ordinateurs datant d'il y a dix ans et où toutes les données sont dans des colonnes dont les noms sont des abréviations dont on a déjà oublié la signification exacte... Ces données sont donc souvent définitivement perdues et ne peuvent plus être utilisées par d'autres et, très souvent, par l'auteur des données lui-même.

Pour pallier ce problème, l'initiative FAIR introduit **le principe “Interopérable (Interopérabilité)”**. Ce principe se résume aux consignes suivantes: pour vos données privilégiez les formats de fichiers et les logiciels non-propriétaires, largement connus et répandus (c'est-à-dire qu'ils ont plus de chances de survivre pendant longtemps que des formats propriétaires, obscurs et/ou peu connus); si vous utilisez les abréviations, les balises, la mise en page ou tout autre type de marquage pour vos données, fournissez des explications détaillées et complètes afin que vos données puissent être comprises et interprétées facilement par toute personne (ou machine).

Sauvegarde / Archivage des données. Il n'est pas rare qu'une fois que l'article ou le projet est terminé, les données utilisées ou produites se

retrouvent éparpillées dans différents fichiers, sur différentes clés USB ou sur les ordinateurs de différents collaborateurs. Par exemple, le texte de la lettre de Léon III à Charlemagne sera dans un fichier "*LettreLéonCh.doc*" sur l'ordinateur du bureau, les images du manuscrit de cette lettre dans un dossier "*Images février 2015 et autres documents*" sur une clé USB rouge qui vous avez prêtée à un ami et l'analyse de la lettre dans un fichier "*A-l-LIII-à-Ch.txt*" sur votre ordinateur portable personnel qui ne s'allume plus depuis mercredi dernier... Le plus souvent il est extrêmement délicat, voire impossible, de restituer ces éléments dans un ensemble cohérent; les données sont donc perdues pour les autres et souvent pour l'auteur des données lui-même.

Pour pallier ce problème, l'initiative FAIR introduit **les principes "Findable (Facile à trouver)" et "Accessible"**. Ces principes se résument aux consignes suivantes: attribuez à chaque jeu de données un identifiant unique et pérenne (par exemple DOI); sauvegardez vos données dans un dépôt institutionnel fiable; grâce aux métadonnées (données sur des données) explicitez ce que vos fichiers contiennent (titre exact, auteur de données, date de production, résumé du contenu, mots-clés, etc.); si votre jeu de données est constitué de plusieurs sous-éléments (textes, images, etc.) reliez ces sous-éléments entre eux (grâce au référencement croisé dans les métadonnées). Suivre ces consignes permet à toute personne et à l'auteur des données lui-même de trouver facilement ces données, comprendre/se rappeler ce que contient chaque fichier et de pouvoir reconstituer l'ensemble des sous-éléments d'un jeu complexe de données.

Droits de la re/utilisation / Autorité des données. Il n'est pas rare qu'une fois que l'article ou le projet est terminé, les données utilisées ou produites ont été transmises par un chercheur à un autre qui les a légèrement modifiées (en corrigeant les erreurs ou en le mettant à jour) puis retransmises à un autre chercheur qui les modifie à son tour ... et qu'entre-temps la législation a changé en rendant une partie des données sensibles ou interdites à la diffusion. L'autorité de ce type de données (qui est auteur de ces données? comment les citer?) et leur droit de l'utilisation (données ouvertes et publiques? données privées? classifiées? sensibles?) ne sont pas toujours très clairs.

Pour pallier ce problème, l'initiative FAIR introduit **le principe "Reusable (Réutilisable)"**. Ce principe se résume aux consignes suivantes: les différents droits/ licences (droit d'auteur, droit d'utilisation, etc.) liés à chaque jeu de données ainsi que la provenance et l'historique des données (qui est l'auteur des données, qui les a collectées, qui les a modifiées, etc.) doivent être clairement mentionnés dans les

métadonnées. Suivre ces consignes permet à toute personne et à l'auteur des données lui-même de savoir s'il est possible ou non, du point de vue légal, de re/utiliser ces données et si elles sont re/utilisables de savoir qui sont ses auteurs et comment les citer.

1.4 Qui FAIRise quoi?

La présentation qui suit fait appel à un écosystème FAIR basé sur un dépôt de données institutionnel (sur les différents écosystèmes FAIR, voir infra "1.5 Des données au dépôt de données: quelques notions principales / Écosystème FAIR").

Les principes FAIR s'appliquent aussi bien aux **métadonnées** qu'aux **données** elles-mêmes (voir *infra* Annexe 1). Le respect des principes qui s'appliquent aux données doit être assuré par **l'auteur des données**, tandis que les principes qui s'appliquent aux métadonnées sont à la charge **du dépôt où les données sont entreposées**.

Le principe "Findable (Facile à trouver)" est atteint avant tout par le biais de la création des **métadonnées** associées au jeu de données et par la mise de ce jeu de données dans un **dépôt de données**. De fait, lors de l'ajout d'un jeu de données dans un dépôt spécialisé, c'est souvent ce dernier qui se charge de la création d'un identifiant unique (PID) et qui propose des formulaires de renseignements à compléter afin de créer des métadonnées associées. Le choix d'un dépôt de données est donc crucial: le dépôt doit fournir un PID valide et unique, le dépôt doit fournir un formulaire le plus complet possible des métadonnées et, enfin, c'est le dépôt qui doit créer les métadonnées facilement reconnaissables et indexables par les moteurs de recherche (en utilisant les langages et les standards largement connus pour les métadonnées, par exemple RDF, OWL, JSON, etc.).

Dès lors, il est possible de dire que c'est avant tout le dépôt où les données sont déposées qui doit répondre au principe "Findable (Facile à trouver)".

Le principe "Accessible" est atteint avant tout grâce au **dépôt où les données sont déposées**. En effet, le dépôt de données doit fournir un moyen (ouvert, gratuit et largement utilisé) d'accéder aux données. En pratique cela signifie qu'une fois que les données sont mises dans un dépôt, ce dernier doit vous fournir un lien (URL) qui permet d'accéder

aux données. Le choix de dépôt de données est donc également crucial pour ce principe; il est indispensable de prendre en compte les moyens que chaque dépôt vous propose pour accéder à vos données. (À noter que le protocole HTTP (=adresse web), bien que le plus connu n'est pas le seul moyen d'accéder aux données digitales. D'autres protocoles existent également, par exemple FTP, SMTP, Microsoft Exchange Server, etc.).

Dès lors, il est possible de dire que c'est avant tout le dépôt où les données sont déposées qui doit répondre au principe "Accessible".

Le principe "Interopérable (Interopérabilité)" est atteint aussi bien par des **métadonnées** que par des **données elles-mêmes**.

En ce qui concerne **les données**, il est indispensable qu'elles soient encodées dans un format largement connu et répandu et qu'elles soient compréhensibles (c'est-à-dire que toutes les abréviations, balises, etc. doivent être explicités). En outre, les données doivent s'appuyer sur un standard/vocabulaire déjà bien documenté (voir par exemple ce recensement des standards répandus: <http://rd-alliance.github.io/metadata-directory>). Ces tâches sont **à la charge de l'auteur des données**.

En ce qui concerne **les métadonnées**, elles doivent également être dans un format largement connu et utilisé (par exemple RDF, OWL, JSON, etc.) et elles aussi doivent être compréhensibles (c'est-à-dire que les métadonnées doivent s'appuyer sur un des standards prévus pour les métadonnées, par exemple DublinCore, Metadata Object Description Schema, etc.). Les métadonnées doivent ainsi être compréhensibles aussi bien aux humains (qui, grâce à ces métadonnées, pourront avoir un premier aperçu des données) qu'aux machines (notamment aux moteurs de recherche qui pourront, grâce à ces métadonnées, indexer les données correctement). Cette tâche de l'interopérabilité des métadonnées **est à la charge de celui qui encode les métadonnées, c'est-à-dire le dépôt de données**.

Dès lors, il est possible de dire que le principe "Interopérabilité" est assuré par l'auteur des données pour les données et par le dépôt où les données sont déposées pour les métadonnées.

Le principe "Reusable (Réutilisable)" est atteint avant tout grâce aux **métadonnées**. Ce principe stipule que certains types d'informations (provenance des données, historiques des changements et des modifications, les droits d'auteurs, les droits d'utilisation, etc.) doivent nécessairement figurer dans les métadonnées associées au jeu de données. Il est indispensable de souligner toutefois que les

métadonnées ne se résument pas aux informations demandées par le principe “Reusable (Réutilisable)”, les métadonnées doivent être “complètes” (voir *supra* le principe “Findable”) et contenir également un ensemble des autres informations (voir *infra* Annexe 4).

*Dès lors, il est possible de dire que ce principe doit être assuré **aussi bien par le dépôt de données** qui doit fournir un formulaire des métadonnées adéquates (c'est-à-dire le formulaire qui contient des champs sur l'historique des données, sur leur provenance, sur leurs droits d'auteurs, etc.) **que par l'auteur des données** qui a l'obligation de remplir ce formulaire.*

En conclusion, on peut dire que le processus de la FAIRisation est assuré par **le dépôt de données** (qui se charge du respect des principes “Findable”, “Accessible”, “Interoperable [pour les métadonnées]” et “Reusable”) et par **l'auteur des données** (qui doit assurer le respect du principe “Interoperable [pour les données]” et remplir tous les champs du formulaire des métadonnées fournies par le dépôt de données).

Il est nécessaire de souligner que le rôle que joue **le dépôt de données** peut être rempli de façon différente. (Sur ce point voir *infra* “1.5 De données au dépôt de données...”. Sur les différentes tâches accomplies par l'auteur des données et le dépôt de données, voir *infra* Annexe 2).

1.5 Des données au dépôt de données: quelques notions principales

Les brèves présentations qui suivent ne sont en aucun cas des définitions arrêtées et elles ne donnent que quelques lectures possibles de notions complexes. L'objectif ici est d'offrir un premier aperçu de ces concepts clés et de proposer quelques repères pour mieux comprendre les motivations et les applications possibles de certains principes FAIR.

Données / Métadonnées

Il est impossible et inutile de dresser dans cette présentation la liste de toutes les définitions des “données” ou des “métadonnées” qui existent. Cependant, dans le cadre de la FAIRisation, il est indispensable de faire

la différence entre les données, les métadonnées et leurs différents types, le modèle de données et l'objet d'étude.

Données vs Métadonnées. Bien que le mot "métadonnées" ne soit apparu que vers la fin des années 1960 dans le contexte de l'arrivée de l'informatique moderne, l'origine du concept est étroitement liée à l'histoire des bibliothèques (pour plus de détail sur ce point, voir *infra* la bibliographie [14]). De fait, le catalogue d'une bibliothèque est la première réalisation pratique de la notion "métadonnée". La bibliothèque est constituée des livres, le catalogue est constitué des fiches qui décrivent les livres présents dans la bibliothèque. Chaque fiche contient, *a minima*, le titre, l'auteur, la date, le lieu de la publication, ainsi qu'un court résumé du livre. L'objectif d'une fiche est de donner au lecteur des indications nécessaires pour trouver un livre dont il a besoin sans qu'il soit nécessaire de consulter le livre lui-même.

La distinction entre un livre et une fiche, ainsi que l'objectif de ladite fiche sont clairs. Dans le contexte des notions "données" et "métadonnées", le livre représente les données (objet à décrire) et la fiche représente les métadonnées (description de l'objet). Si on transpose ces concepts dans un monde digital, la distinction entre les données et les métadonnées doit également être aisée. De fait, un document numérique qui contient un texte ou une image représente une donnée, et une "fiche digitale" qui décrit le contenu de ce document (titre, auteur, date de production, résumé...) représente les métadonnées. *Les données sont dès lors les objets que l'on décrit et les métadonnées sont les descriptions de ces objets.* L'objectif des métadonnées est de permettre aux utilisateurs (humains) ainsi qu'aux ordinateurs et moteurs de recherche (machines) d'identifier et de trouver un document numérique en fonction des critères spécifiques attribués à ce document (titre, auteur, date, etc.) sans qu'il soit nécessaire de consulter le document lui-même.

Différents types de métadonnées. Si le titre, l'auteur ou la date de publication d'un document, qu'il soit numérique ou non, peuvent être utiles jusqu'à un certain point, ces renseignements ne seront pas suffisants pour quelqu'un qui cherche un contenu spécifique, par exemple un document qui ne contient que du texte en latin. Et si on ne recherche que des documents écrits à Rome? Il se pose alors la question de savoir quels renseignements doivent contenir les métadonnées. Doivent-elles être exhaustives? Faut-il y mettre toutes les informations potentiellement utiles?

Il est rarement possible de véhiculer toute la complexité et la richesse d'un objet à travers sa description. Chaque description est résolument incomplète et ne peut représenter qu'un des aspects de l'objet qu'elle décrit. Il sera, par exemple, peu judicieux d'essayer de mettre sur une seule carte de France toutes les informations que l'on dispose sur ce pays; il est plus commode de recourir à des cartes différentes pour chaque type d'information que l'on veut représenter (carte administrative, topographique, routière, etc.).

Le même principe s'applique aux métadonnées. De fait, les métadonnées peuvent être de différents types où chaque type n'exprime qu'un des aspects de l'objet que l'on décrit (pour plus de détail sur ce point, voir *infra* la bibliographie [14]). Par exemple, pour la même photo numérique un informaticien voudra savoir la taille de la photo en mégaoctets et sa résolution en DPI, un photographe souhaitera plutôt connaître l'appareil et les filtres utilisés pour l'acquisition de la photo et un historien de l'art s'intéressera avant tout à l'histoire de la peinture représentée sur la photo. Ou bien encore, un fichier xml qui contient une charte du XII^e siècle peut être accompagné des métadonnées générales (titre, auteur, date de production, langue, etc.), des métadonnées davantage centrées sur l'aspect informatique (taille de fichier, format d'encodage, standard, etc.) et des métadonnées propres à la recherche historique (sources manuscrites, bibliographie, éditions, type, état et qualité du support, etc.).

Les métadonnées de différents types offrent alors des renseignements différents sur le même objet. Ces types peuvent varier aussi bien en fonction de notre problématique que de l'objet lui-même que l'on décrit. On utilisera, par exemple, les différentes métadonnées pour décrire un texte latin du Moyen Âge et une peinture rupestre du Paléolithique supérieur. Les principes FAIR eux-mêmes incitent à introduire des métadonnées spécifiques (voir *infra* Annexe 4). Dès lors, *par un type de métadonnées on entend des renseignements sur l'objet que l'on décrit (=donnée) qui peuvent être regroupés dans un ensemble spécifique.*

De nombreux types de métadonnées ont été étudiés et ont fait l'objet d'une standardisation. Par exemple, les métadonnées que l'on peut appeler "essentiels" (titre, auteur, date de la publication, langue, etc.) ont donné lieu à un standard de métadonnées "Dublin Core" ("core" est traduit alors comme "essentiel") qui comprend 15 éléments basiques (qui peuvent, si nécessaire, être enrichis par quelque 40 "termes"). Les métadonnées plus spécifiques ont été standardisées au sein de leurs domaines respectifs, par exemple le standard "Darwin Core" pour les métadonnées biologiques ou bien encore "Metadata

Object Description Schema" (MODS) pour des métadonnées bibliographiques.

Cependant, il ne sera guère envisageable de prévoir un standard pour chaque type de métadonnées. Il existe dès lors la possibilité de mettre en place ses propres standards (appelés alors "schéma") pour décrire un type de métadonnées personnelles spécifiques. Les formats sous lesquels les métadonnées peuvent être encodées sont également très variés. À l'heure actuelle, il existe plusieurs dizaines de formats et de standards différents pour encoder les métadonnées et leur choix est souvent un point important dans l'organisation des données d'un projet.

(En théorie, rien n'empêche d'associer plusieurs fichiers de métadonnées au même fichier de données. Dans ce cas, chaque fichier de métadonnées peut s'appuyer sur des formats et des standards des métadonnées différents. Or, il est indispensable de se rappeler qu'une telle configuration nécessite que tous ces fichiers soient reliés entre eux, par exemple par un référencement interne croisé de leurs identifiants uniques. Quelle que soit la configuration mise en place, elle dépendra nécessairement des besoins et des ressources disponibles de chaque projet. Sur l'emplacement des métadonnées voir également infra "Données et Métadonnées, ensembles ou séparées?")

(Pour une présentation plus détaillée des différents langages, formats ou standards des métadonnées voir la bibliographie infra [6], [14]).

Modèle des données vs Objet d'étude. Malgré le fait que la distinction entre les données et les métadonnées décrites plus haut soit d'apparence claire, dans certains contextes la séparation entre les deux peut être plus délicate. C'est le cas, par exemple, d'un acte médiéval quand son contenu (=donnée) est accompagné par des renseignements supplémentaires (=métadonnées) aussi bien sur le langage et la forme de discours que sur le support du document, l'encre et l'écriture. Toutes ces précisions peuvent être aussi une source riche d'informations pour un historien en devenant donc des "données".

Ce même processus peut être observé à l'exemple, déjà cité, des livres et des fiches. De fait, pour quelqu'un qui s'intéresse aux livres, ces derniers sont des "données" et les fiches de catalogue qui les décrivent sont des "métadonnées". En revanche, pour quelqu'un qui s'intéresse à l'histoire des catalogues des bibliothèques, ce sont les fiches elles-mêmes qui deviennent des "données" à étudier... Il est donc

indispensable de se rappeler que les métadonnées des uns peuvent être les données des autres.

Pour mieux saisir ce processus d'interchangement entre les données et les métadonnées, il est nécessaire de faire la différence entre un modèle conceptuel et un modèle logique de données d'un côté et l'objet d'étude de l'autre. **Un modèle conceptuel (ou formel) de données** désigne le concept (ou objet formel) à décrire (ici, par exemple, le texte d'un acte) et tous les renseignements supplémentaires (ici, auteur de l'acte, date de l'écriture, lieu d'écriture, support de l'écriture, etc.) qui peuvent accompagner le concept que l'on décrit. **Un modèle logique de données**, de son côté, décrit comment (entités, attributs, relations, etc.) il est prévu de mettre en pratique le modèle conceptuel. Par exemple, cela peut être un fichier (ou une table, s'il s'agit de la base de données) avec le texte de l'acte, accompagné des informations auxiliaires sur cet acte (auteur de l'acte, date de l'écriture, lieu d'écriture, support de l'écriture, etc.) écrites soit dans le même fichier (ou la même table) soit dans un fichier (ou une table) joint. Enfin, toute information, qu'elle soit auxiliaire ou non, contenue dans les modèles de données peut être un **objet d'étude** qui est défini alors par la problématique de la recherche. "L'objet d'étude" d'une recherche, dès lors, n'est pas la même chose que "l'objet formel" qui est la base d'un modèle des données. Si les modèles sont assez fixes (il est peu probable que la base de données soit reconstruite ou les fichiers re-encodés au cours d'un projet), l'objet d'étude est plus "mouvant" et peut changer au gré des questions posées.

La "métadonnée" n'est donc qu'un rôle que n'importe quelle donnée peut jouer à l'égard des autres données. De ce point de vue, la question de savoir où sont les données et où sont les métadonnées peut paraître erronée. Il est plus opportun de se demander: "Dans le contexte de ces données prises comme un objet à décrire, quelles sont les autres données qui joueront le rôle des métadonnées?". La réponse dépend inévitablement du contexte dans lequel cette question est posée (s'agit-il du modèle conceptuel ou logique ou de l'objet d'étude?) et de quel est l'auteur de cette question (administrateur, informaticien, scientifique...).

(Pour plus de détails, voir infra la bibliographie [6], [8], [9], [10], [14]).

Données et Métadonnées, ensemble ou séparées? Un des derniers points concerne l'emplacement des métadonnées. Deux solutions sont possibles: stocker les métadonnées dans le même endroit que les données (par exemple le même fichier ou la même table de la base de données) ou séparément. Le choix dépend de l'architecture et de

l'organisation des données que vous souhaitez ou pouvez mettre en place.

Il faut toutefois garder à l'esprit que l'objectif des métadonnées est de permettre de trouver, rapidement et facilement, un objet spécifique (voir *supra* "Données vs Métadonnées"). Placer les métadonnées dans un fichier séparé donne la possibilité - aussi bien à l'utilisateur (humain) qu'à l'ordinateur ou au moteur de recherche (machine) - de parcourir des fichiers courts et légers avec des informations essentielles (=métadonnées) au lieu de traiter des documents volumineux et complexes (=données) pour y chercher des renseignements particuliers. Enfin, il faut se rappeler que si les métadonnées sont stockées séparément des données, elles [données et métadonnées] doivent être reliées par un référencement croisé (par exemple un PID).

(Sur ce point voir *infra* la bibliographie [6], [8], [14], cf. également Annexe 1, le principe F1 et surtout le principe F3)

Jeu de données

Les données auxquelles les principes FAIR sont appliqués peuvent être désignées comme *Objet Digital FAIR* ou FDO (FAIR Digital Object) [7], [15], [17]. Cela signifie que dans le cadre de la FAIRisation des données, l'identifiant pérenne et les métadonnées sont associés à un FDO [1].

Les données scientifiques, notamment en histoire, se présentent le plus souvent sous la forme d'un corpus constitué de plusieurs documents. Il se pose alors la question de savoir si le FDO correspond à un corpus entier ou à un seul document. Les recommandations FAIR sont muettes à ce sujet. La bibliographie précise seulement que le FDO doit être "*une entité significative*" pour la FAIRisation [1], [7], [16]. À la fin, le choix dépend de l'écosystème FAIR choisi, soit une infrastructure personnelle soit un dépôt de données (sur les différents types d'écosystème FAIR voir *infra* "Écosystème FAIR").

Si c'est l'infrastructure personnelle qui a été retenue, il appartient à chaque projet de choisir le niveau de la graduation préférable. Il est possible de proposer l'accès aux données aussi bien document par document qu'au corpus entier (en tant qu'un fichier unique) ou bien les deux.

En revanche, dans le cas du choix d'un dépôt institutionnel comme une infrastructure principale de la FAIRisation, il semble plus judicieux d'opter pour le jeu de toutes les données d'un corpus et non pour une seule valeur atomique (c'est à dire notice, acte, charte, lettre, etc.) comme un FDO. Ce choix est dicté aussi bien par les pratiques déjà existantes (les dépôts Zenodo et Harvard Dataverse, par exemple, ne

semblent héberger que les “dataset / jeu de données”) que par une plus grande simplicité de cette approche (déposer un seul fichier du corpus plutôt que plusieurs centaines de documents un par un).

Quel que soit le procédé utilisé, il est important de préciser ce qui est entendu par le “jeu de données” dans le contexte de la FAIRisation. Un “jeu de données” désigne l'ensemble de toutes les données du corpus réuni dans un seul fichier (ou archive numérique, par exemple sous format zip). Il est important de se rappeler qu'il peut exister plusieurs jeux de données où chaque “jeu” réunit des données analogues, par exemple les images, les textes, les notes, etc. Par la suite, ces différents ensembles pourront être reliés grâce aux référencements croisés des identifiants uniques dans les métadonnées (principe FAIR I3).

(Pour plus d'informations sur le FDO voir infra la bibliographie, notamment [7], [15]).

Écosystème FAIR

La mise en place des principes FAIR passe par plusieurs procédures et peut s'appuyer sur des outils différents. On parle alors d'un écosystème FAIR qui comprend des acteurs, services et concepts très variés (identifiant pérenne, FAIR Digital Object, Plan de Gestion des Données, dépôt des données, etc.). Bien qu'il revienne à chaque projet de choisir les moyens par lesquels il souhaite atteindre les différents principes FAIR, il est possible d'envisager deux types majeurs d'écosystème.

Le premier type s'appuie sur un dépôt de données institutionnel¹ (**Figure 1**). Dans le cas où il est préférable de se concentrer sur la récolte et l'analyse des données, il est judicieux de confier toutes les autres opérations, comme le stockage, l'archivage, la diffusion et, même, l'affichage des données brutes, à un dépôt déjà existant.

¹ Le terme “repository”, utilisé en anglais pour désigner un “dépôt” de données, peut, parfois, être traduit en français comme “référentiel”. Voir par exemple “[Wikipedia-Dépôt \(informatique\)](#)”. Cependant un “référentiel” est une notion polysémique qui désigne, dans la langue française, avant tout un système de référence plutôt qu'un dépôt. L'utilisation de ce vocable peut donc prêter à confusion et il est indispensable d'y prendre garde. De fait, la notion “data repository”, qui désigne alors “un endroit où les données sont déposées”, traduite en français par “un référentiel des données”, peut être comprise, à tort, comme “un système de référence des données”. C'est donc la traduction “dépôt de données” qui doit être systématiquement privilégiée.

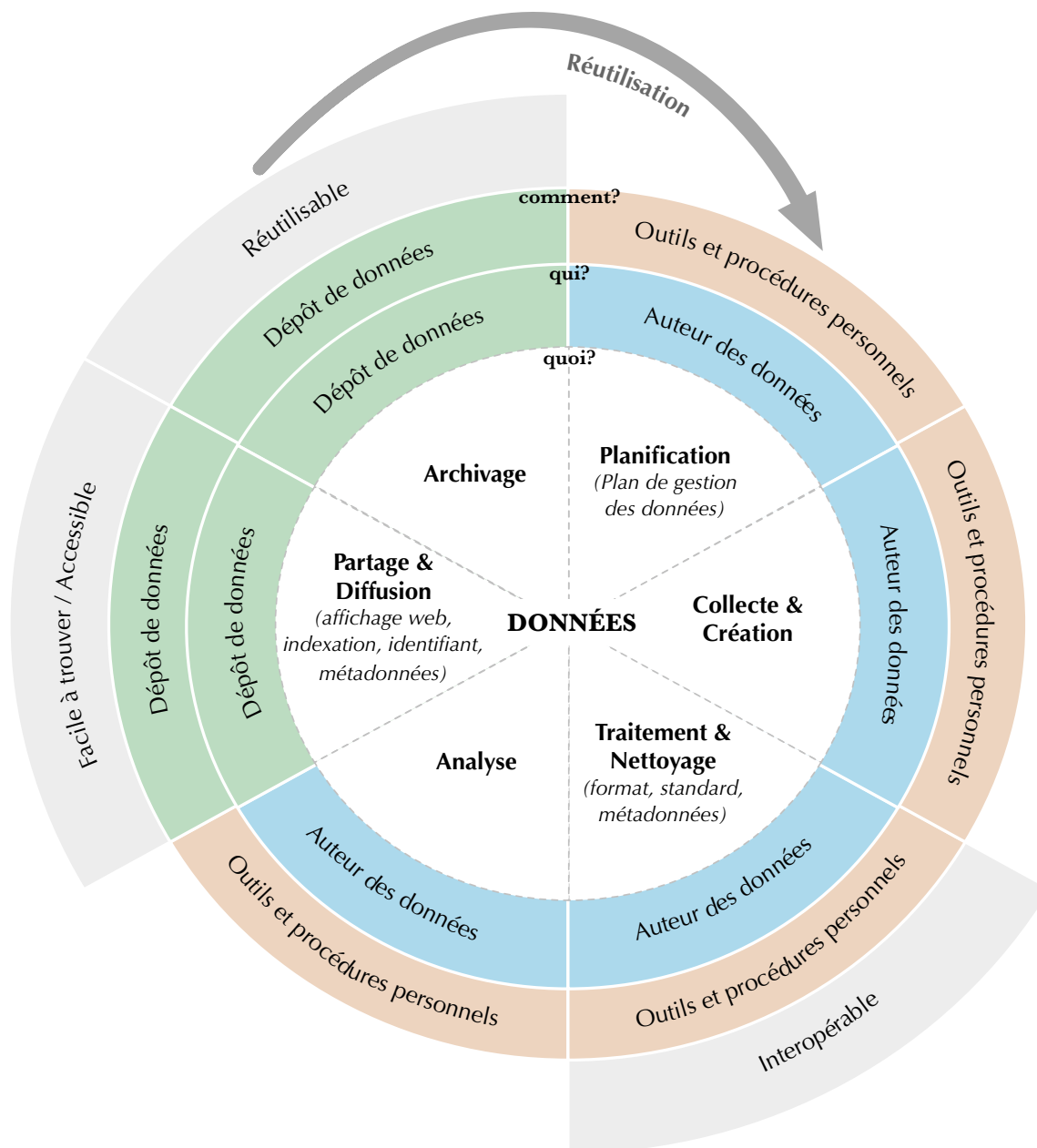


Figure 1. Principes FAIR et écosystème basé sur un dépôt institutionnel dans le contexte du cycle de vie des données.

De fait, il est indispensable de se rappeler que le dépôt de données ne remplit pas uniquement la fonction de l'archivage des données. La majorité des dépôts institutionnels s'occupe de tâches très diverses comme l'attribution des identifiants, la gestion des métadonnées, l'indexation et la diffusion des données auprès des moteurs de recherche, le versioning, la mise en place de l'accès aux données, etc. (À noter toutefois que d'ordinaire c'est un jeu de toutes les données

d'un corpus qui est versé dans un dépôt. Sur ce point, voir *supra* "Jeu de donnée"). Dans ce type d'écosystème, l'auteur des données ne s'occupe que de l'encodage des données dans un format et avec un standard connus et répandus (principes FAIR I1 et I2). (Le programme de la FAIRisation proposé ci-dessus s'appuie sur cette approche, voir *infra* "2 Étapes de la FAIRisation").

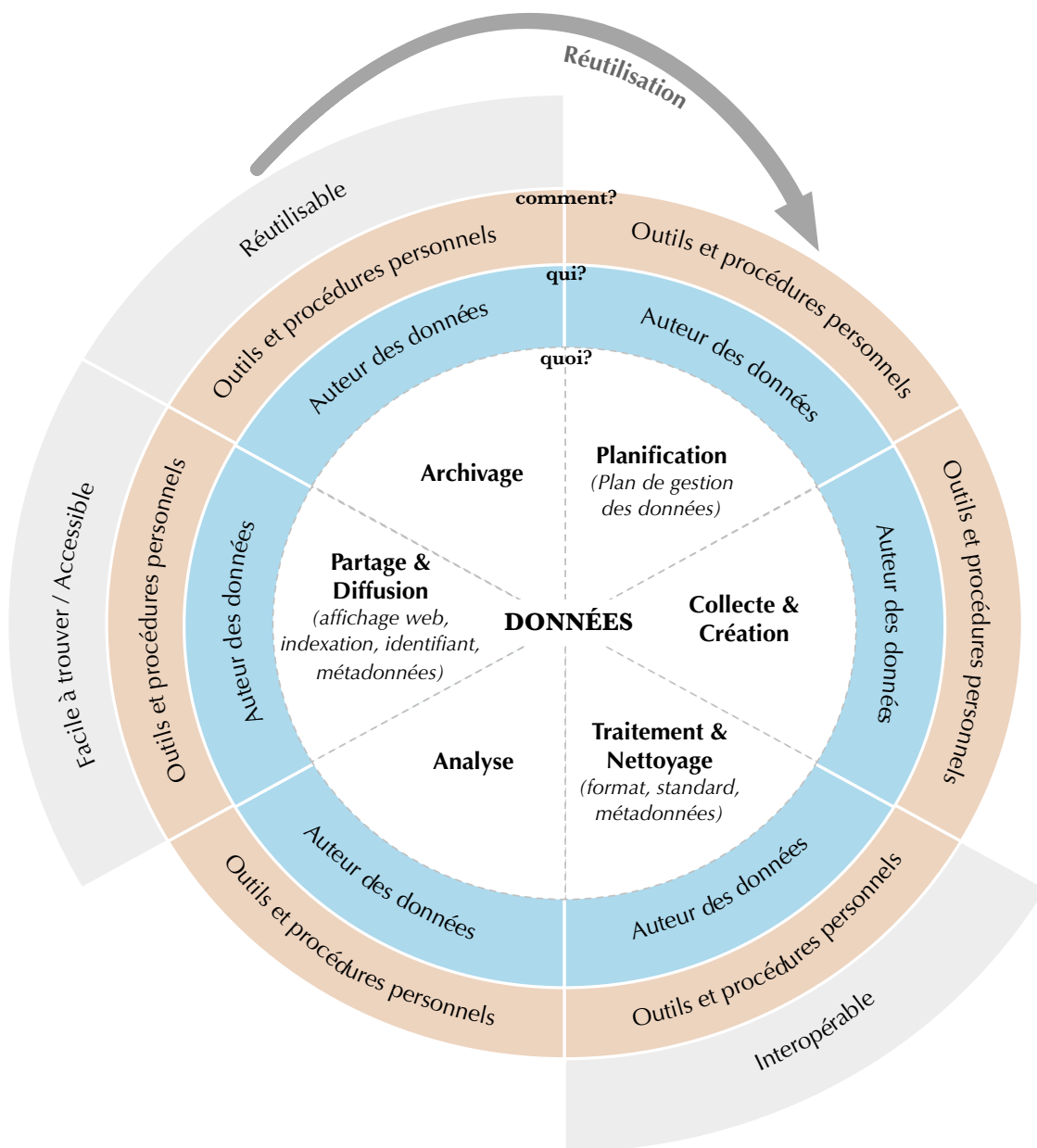


Figure 2. Principes FAIR et écosystème basé sur des outils personnels dans le contexte du cycle de vie des données.

Le deuxième type d'écosystème s'articule principalement autour d'une infrastructure personnelle (**Figure 2**). Cette approche est judicieuse s'il

est nécessaire d'avoir la main mise sur chaque étape du cycle de vie des données, de leur récolte à leur diffusion. L'infrastructure personnelle peut se matérialiser par la création de ses propres outils de gestion des métadonnées (par exemple à l'aide de *FAIR Data Point*, voir [9]), des sites web ou des plateformes personnels pour l'affichage des données, des serveurs locaux pour leur stockage, etc. (Voir par exemple l'infrastructure FAIR mis en place par le consortium MASA: <https://masa.hypotheses.org/files/2021/02/MASAecosystem2020-scaled.jpg>). Dans le cadre de cette deuxième approche, le recours à un dépôt de données peut également trouver sa place en tant qu'un des éléments de l'infrastructure (par exemple uniquement pour le partage et l'exposition des données).

Bien que, dans une grande partie des cas, l'écosystème, qui s'appuie grandement sur l'utilisation de dépôt de données, soit le moyen le plus simple et le plus rapide pour satisfaire les principes FAIR, il revient à chaque projet de choisir l'infrastructure qu'il estime le plus adaptée à ses besoins et à ses ressources. À la fin, quelle que soit l'infrastructure choisie, il est indispensable de s'assurer qu'elle réponde à tous les principes FAIR attendus (voir *infra* Annexe 2 et Annexe 3).

(Pour plus de détails sur la notion de l'écosystème FAIR voir la bibliographie [10], [12], [16], [17], [18]).

2 Étapes de la FAIRisation

Les étapes suivantes décrivent le processus complet de la FAIRisation qui peut être appliqué à un jeu de données d'un corpus afin que ce dernier réponde à tous les critères FAIR. Ce programme de la FAIRisation fait appel à un écosystème FAIR basé sur un dépôt des données institutionnel (sur les différents écosystèmes FAIR, voir *supra* "1.5 Des données au dépôt de données: quelques notions principales / Écosystème FAIR").

1. Vérification des données

1.1. S'assurer que les langages et les formats/extensions utilisés pour l'encodage des données sont connus, largement utilisés et, de préférence, libres. (Il est impossible de définir in fine ce qui est un format/langage "connu" ou "largement utilisé". On s'appuyer donc sur le bon sens... Privilégiez toujours les formats libres, par exemple csv (format libre) au lieu de xls (format propriétaire de Microsoft), txt (format libre) au lieu de doc (format propriétaire de Microsoft), etc. Pour plus d'informations sur le choix des formats à privilégier pour la pérennité des données de la recherche voir ce recensement: <https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats>). Si cela n'est pas le cas, expliquer clairement ce qui a dicté le choix d'un langage ou d'un format particulier et indiquer tous les logiciels (nom, version, système d'exploitation) et outils nécessaires pour la lecture des données. (Ces informations seront par la suite utilisées dans les métadonnées).
FAIR: I1

1.2. S'assurer que les données sont explicites et compréhensibles en l'état. Cela signifie que toutes les abréviations, tous les termes ambigus ou toutes les balises utilisées dans la description des données doivent être clairement expliqués ou avoir déjà une documentation existante. (Pour plus d'informations sur les différents standards voir ce recensement: <http://rd-alliance.github.io/metadata-directory/standards/>).

Trois options sont possibles: i) les données sont explicites en l'état; ii) les données ne sont pas explicites et sont accompagnées d'une documentation détaillée personnelle; iii) les données ne sont pas explicites et sont accompagnées d'un renvoi vers une documentation

détaillée déjà existante (par exemple TEI P5 TEI Guidelines). (Ces informations seront par la suite utilisées dans les métadonnées). **FAIR I2**

2. Préparation des métadonnées

2.1. Préparation des métadonnées complètes.

2.1.1. S'assurer que les métadonnées contiennent le plus de détails possible. (Se référer au modèle des métadonnées détaillées, voir Annexe 4) **FAIR: F2**

2.1.2. S'assurer que les données disposent une licence et qu'elle est clairement mentionnée dans les métadonnées. (Les données peuvent avoir tous types de licence, y compris les licences "non libres de droits"). **FAIR: R1.1**

2.1.3. S'assurer que les métadonnées contiennent un historique détaillé de la provenance des données. (Par exemple, les données peuvent être produites durant un autre projet et ensuite réutilisées, corrigées et modifiées par le présent projet). **FAIR: R1.2**

2.1.4. S'assurer que les données utilisent des standards (s'il en existe) de présentation, d'organisation ou d'archivage communément utilisés dans le domaine scientifique auquel ces données appartiennent. (Ces informations seront par la suite utilisées dans les métadonnées).

(Bien qu'un "standard" commun ne semble pas exister en sciences humaines et sociales, certains projets ou programmes de recherche peuvent s'appuyer sur des conventions similaires. Par exemple, l'organisation des données dans le même type de tableau, l'utilisation des mêmes principes de balisage, etc. Si c'est le cas, ces conventions doivent être mentionnées, par exemple dans les métadonnées. Ne pas confondre avec I2. Le I2 vise à rendre les données compréhensibles / interopérables grâce à la documentation associée à ces données. Le R1.3 vise à faciliter la réutilisation des données grâce aux indications générales sur les standards et les conventions sur lesquels les données s'appuient. Pour plus d'informations sur les différents standards voir ce recensement: <http://rd-alliance.github.io/metadata-directory/standards/>) **FAIR: R1.3**

2.2. Relier les données.

S'assurer que dans les cas où les présentes données peuvent être reliées à d'autres données (par exemple, les données textuelles peuvent être reliées aux images des manuscrits), les métadonnées contiennent des références à ces données reliées (par exemple le PID des données reliées, voir F1). Le type de lien entre les données doit être explicitement mentionné. (Par exemple, "ce texte est la transcription de ce manuscrit" ou "ce texte a été écrit par cette personne"). **FAIR: I3.**

3. Mise des données dans un dépôt

3.1. Enregistrement des données et des métadonnées.

3.1.1. Choisir une ressource consultable (par exemple, un dépôt de données) qui répond à tous les critères FAIR applicables aux métadonnées (F1, F3, F4, A1.1, A1.2, A2, I1, I2, I3, R1.1, R1.2, R1.3). La plupart des grands dépôts de données affichent de façon claire leur conformité aux critères FAIR. *(Pour plus de détails sur le choix de dépôt et sur les critères FAIR auxquels il doit répondre voir infra Annexe 3)*

3.1.2. Enregistrer les données et les métadonnées y associées dans une ressource consultable choisie précédemment. **FAIR: F4**

3.2. Vérification de l'accessibilité et de l'exactitude des (méta)données.

3.2.1. Vérifier que les données possèdent un identifiant persistant et unique (PID) (fourni par la ressource où les données sont enregistrées). **FAIR: F1**

3.2.2. Vérifier que le PID des données est mentionné dans les métadonnées. **FAIR: F3**

3.2.3. Vérifier que les données sont accessibles via un protocole de communication ouvert, gratuit et largement utilisé. Si les données sont enregistrées dans une ressource qui est accessible via un site internet, ce critère est automatiquement rempli. *(Par un protocole de communication "ouvert, gratuit et largement utilisé", on entend avant tout un protocole informatique libre et connu comme HTTP/HTTPS (=site*

web), FTP etc. Les autres “protocoles de communications” qui s'appuient sur des logiciels propriétaires et non universellement implantés (par exemple Skype, Microsoft Exchange Server, etc.) ne répondent pas à ce critère). **FAIR: A1.1**

3.2.4. Vérifier que les données sont accessibles via un protocole de communication qui permet, si nécessaire, de mettre en place une procédure d'authentification et d'autorisation. (FAIR data ne signifie pas OpenData. Le protocole via lequel les données et les métadonnées sont accessibles, doit avoir la possibilité, en cas de besoin, de restreindre l'accès aux données, par exemple uniquement aux utilisateurs possédant un login et un mot de passe). **FAIR: A1.2**

3.2.5. Vérifier que les métadonnées seront accessibles même lorsque les données elles-mêmes ne seront plus disponibles. (De façon ordinaire ce critère est considéré comme rempli, si la condition F4 est satisfaite). **FAIR: A2**

Bibliographie

- [1] *A Persistent Identifier (PID) policy for the European Open Science Cloud*, European Commission, Directorate-General for Research and Innovation, 2020. doi: 10.2777/926037
- [2] Bloemers M, Montesanti A. "The FAIR funding model: Providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices", *Data Intelligence* 2(2020), 171–180. doi: 10.1162/dint_a_00039
- [3] David R, et al. "FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles", *Data Science Journal*. 2020;19(1):32. doi: 10.5334/dsj-2020-032
- [4] *FAIR Data Maturity Model: specification and guidelines*, RDA FAIR Data Maturity Model Working Group, 2020. doi: 10.15497/rda00045
- [5] *Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, 2016. En ligne: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- [6] *Handbook of Metadata, Semantics and Ontologies*, éd. Sicilia M.-A., World Scientific Publishing, 2014. doi: 10.1142/7077
- [7] Harjes J. et al. "FAIR digital objects in environmental and life sciences should comprise workflow operation design data and method information for repeatability of study setups and reproducibility of results", *Database: the journal of biological databases and curation* (2020). doi: 10.1093/database/baaa059
- [8] Hay D. C., *Data model patterns: a metadata map*, Elsevier, 2006. doi: 10.1016/B978-0-12-088798-9.X5000-1
- [9] Jacobsen A. et al. "A Generic Workflow for the Data FAIRification Process." *Data Intelligence* 2 (2020): 56-65. doi: 10.1162/dint_a_00028
- [10] Jacobsen A. et al. "FAIR principles: Interpretations and implementation considerations", *Data Intelligence*. 2020;2 (1-2) :10-29. doi: 10.1162/dint_r_00024
- [11] Landi A. et al. "The "A" of FAIR – as open as possible, as closed as necessary", *Data Intelligence* 2(2020), 47–55. doi: 10.1162/dint_a_00027

- [12] de Miranda Azevedo R., Dumontier M., "Considerations for the Conduction and Interpretation of FAIRness Evaluations", *Data Intelligence*. 2020 Jan 1;2(1-2):285-292. doi: 10.1162/dint_a_00051
- [13] Pergl R., et al. 2019. "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning", *Data Science Journal*, 18: 59, pp. 1–8. doi: 10.5334/dsj-2019-059
- [14] Pomerantz J., *Metadata*, MIT Press, 2015 doi: 10.7551/mitpress/10237.001.0001
- [15] Schwardmann U., "Digital Objects – FAIR Digital Objects: Which Services Are Required?", *Data Science Journal*, 19(1) 2020. doi: 10.5334/dsj-2020-015
- [16] Thompson M., et al. "Making FAIR Easy with FAIR Tools: From Creolization to Convergence." *Data Intelligence* 2 (2020): 87-95. doi: 10.1162/dint_a_00031
- [17] *Turning FAIR into reality. Final report and action plan from the European Commission expert group on FAIR data*. Luxembourg: Publications Office of the European Union, 2018. doi: 10.2777/1524
- [18] Wilkinson M., et al. "The FAIR Guiding Principles for scientific data management and stewardship", *Science Data* 3, 160018 (2016). doi: 10.1038/sdata.2016.18
- [19] Wilkinson M., et al. "A design framework and exemplar metrics for FAIRness", *Science Data* 5, (2018), 1–4. doi: 10.1038/sdata.2018.118
- [20] Wise J., et al. "Implementation and relevance of FAIR data principles in biopharmaceutical R&D", *Drug Discovery Today*, Volume 24/4 April 2019. doi: 10.1016/j.drudis.2019.01.008

Annexe 1

Principes FAIR

D'après "GO FAIR: FAIR Principles". En ligne: <https://www.go-fair.org/fair-principles>. Le nombre total des principes est de 15.

Rappel: Tous les principes (sauf F2, F3, A2) s'appliquent aussi bien aux données qu'aux métadonnées; cela explique l'utilisation du vocable "(méta)données" (pour les principes FAIR étendus voir *infra* Annexe 2). La partie "données" est garantie par l'auteur des données, la partie "métadonnées" est garantie par le dépôt de données (voir *supra* "1.4 Qui FAIRise quoi?" et "1.5 De données au dépôt de données: quelques notions principales / Écosystème FAIR", voir également *infra* Annexe 2 et Annexe 3).

Findable (= Facile à trouver)

F1. Les (méta)données possèdent un identifiant persistant et unique au monde (*persistent identifier*, PID).

F2. Les données sont décrites avec des métadonnées riches.

F3. Les métadonnées incluent, d'une façon claire et explicite, l'identifiant des données qu'elles décrivent.

F4. Les (méta)données sont enregistrées ou indexées dans une ressource consultable.

Accessible

A1. Les (méta)données sont récupérables par leur identifiant en utilisant un protocole standard de communication.

A1.1. Le protocole est ouvert, gratuit et largement utilisé.

A1.2. Le protocole permet, si nécessaire, une procédure d'authentification et d'autorisation.

A2. Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.

Interoperable (= Interopérable)

I1. Les (méta)données utilisent un langage formel, accessible, partagé et largement répandu pour la représentation des connaissances.

I2. Les (méta)données utilisent des vocabulaires qui suivent les principes FAIR.

I3. Les (méta)données incluent des références qualifiées à d'autres (méta)données.

Reusable (= Réutilisable)

R1. Les méta(données) sont décrites d'une façon complète avec plusieurs attributs précis et pertinents.

R1.1. Les (méta)données sont publiées avec une licence d'utilisation des données claire et accessible.

R1.2. Les (méta)données disposent d'une provenance détaillée.

R1.3. Les (méta)données répondent aux standards communautaires de leur domaine.

Annexe 2

Principes FAIR étendus et les différents acteurs

Rappel: Les principes FAIR étendus désignent chaque principe FAIR doublé dans le cas s'il s'applique aussi bien aux données qu'aux métadonnées. Le nombre total des principes FAIR est de 15 (voir *supra* Annexe 1). Le nombre des principes FAIR étendus est de 27.

Chaque principe FAIR peut être accompli par des acteurs différents (*sur les différents écosystèmes FAIR voir supra "1.5 Des données au dépôt de données: quelques notions principales / Écosystème FAIR"*). Voir, par exemple, comment les différents dépôts remplissent les principes FAIR: le dépôt Zenodo <https://about.zenodo.org/principles/>, le dépôt Dataverse <https://dataverse.org/software-features>.

Principes FAIR étendus	Auteur des données	Dépôt de données ou Infrastructure personnelle
F1. Les données possèdent un identifiant persistant et unique au monde (persistant identifier, PID).		X
F1. Les métadonnées possèdent un identifiant persistant et unique au monde (persistant identifier, PID).		X
F2. Les données sont décrites avec des métadonnées riches.	X (remplir le formulaire des métadonnées)	X (fournir le formulaire des métadonnées)
F3. Les métadonnées incluent, d'une façon claire et explicite, l'identifiant des données qu'elles décrivent.		X
F4. Les données sont enregistrées ou indexées dans une ressource consultable.		X
F4. Les métadonnées sont enregistrées ou indexées dans une ressource consultable.		X
A1. Les données sont récupérables par leur identifiant en utilisant un protocole standard de communication.		X

A1. Les métadonnées sont récupérables par leur identifiant en utilisant un protocole standard de communication.		X
A1.1. Le protocole est ouvert, gratuit et largement utilisé. [pour les données]		X
A1.1. Le protocole est ouvert, gratuit et largement utilisé. [pour les métadonnées]		X
A1.2. Le protocole permet, si nécessaire, une procédure d'authentification et d'autorisation. [pour les données]		X
A1.2. Le protocole permet, si nécessaire, une procédure d'authentification et d'autorisation. [pour les métadonnées]		X
A2. Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.		X
I1. Les données utilisent un langage formel, accessible, partagé et largement répandu pour la représentation des connaissances.	X	
I1. Les métadonnées utilisent un langage formel, accessible, partagé et largement répandu pour la représentation des connaissances.		X
I2. Les données utilisent des vocabulaires qui suivent les principes FAIR.	X	
I2. Les métadonnées utilisent des vocabulaires qui suivent les principes FAIR.		X
I3. Les données incluent des références qualifiées à d'autres données .	X (remplir le champ approprié des métadonnées)	X (fournir le champ approprié des métadonnées)
I3. Les métadonnées incluent des références qualifiées à d'autres métadonnées .	X (remplir le champ approprié des métadonnées)	X (fournir le champ approprié des métadonnées)
R1. Les données sont décrites d'une façon complète avec plusieurs attributs précis et pertinents.	X (remplir le formulaire des métadonnées)	X (fournir le formulaire des métadonnées)
R1. Les métadonnées sont décrites d'une façon complète avec plusieurs attributs précis et pertinents.		X

R1.1. Les données sont publiées avec une licence d'utilisation des données claire et accessible.	X (remplir le champ approprié des métadonnées)	X (fournir le champ approprié des métadonnées)
R1.1. Les métadonnées sont publiées avec une licence d'utilisation des données claire et accessible.		X
R1.2. Les données disposent d'une provenance détaillée.	X (remplir le champ approprié des métadonnées)	X (fournir le champ approprié des métadonnées)
R1.2. Les métadonnées disposent d'une provenance détaillée.		X
R1.3. Les données répondent aux standards communautaires de leur domaine.	X (remplir le champ approprié des métadonnées)	X (fournir le champ approprié des métadonnées)
R1.3. Les métadonnées répondent aux standards communautaires de leur domaine.		X

Annexe 3

Choix d'un dépôt de données

Rappel: Un dépôt de données doit garantir tous les principes FAIR applicables aux métadonnées (c'est-à-dire, tous les principes sauf F2, voir *supra* "1.4 Qui FAIRise quoi?" et Annexe 1).

Comment choisir un dépôt de données?

- choisir le dépôt le plus utilisé dans votre discipline;
- ou sinon, choisir le dépôt le plus utilisé dans votre institution;
- ou sinon, choisir un dépôt "généraliste".

Où chercher un dépôt de données pour votre discipline?

- *Registry of Research Data Repositories*: <https://www.re3data.org>
- *OpenDOAR*: <https://v2.sherpa.ac.uk/opensoar/search.html>
- *DATAACC. Déposer ses données en ligne: où et comment*: <https://www.dataacc.org/vos-besoins/valoriser-ses-donnees/deposer-ses-donnees-en-ligne-ou-et-comment>

Les dépôts des données "généralistes" les plus utilisés:

- *Zenodo (CERN Data Centre / EU Horizon 2020 / OpenAIRE)*: <https://zenodo.org>
- *Harvard Dataverse*: <https://dataverse.harvard.edu>
- *Figshare*: <https://figshare.com>
- *Mendeley Data*: <https://data.mendeley.com>
- *Open Science Framework (OSF)*: <https://osf.io>
- *Dryad*: <https://datadryad.org/stash>

Annexe 4

Modèle (prévisionnel) des métadonnées

Rappel: Il est utile de rappeler que les principes FAIR ne sont que des recommandations qui visent à rendre les données scientifiques plus accessibles et plus compréhensibles et leurs sauvegarde et réutilisation plus pérennes. La FAIRitude [FAIRness] complète des données n'est cependant qu'un objectif théorique difficilement atteignable (sur ce point, voir la bibliographie [2], [3], [10], [12]). Par conséquent, chaque champ des métadonnées doit être rempli le plus complètement possible, mais dans des limites raisonnables.

Champ	Correspondance éléments Dublin Core	Correspondance champs Zenodo
Identifiant (PID): <i>(L'identifiant est attribué automatiquement par le dépôt lors de la mise en ligne.)</i>	DC.Identifier	Digital Object Identifier ⁺
Date de la mise en ligne (des données):	DC.Date	Publication date ⁺
Intitulé (du jeu de données du corpus):	DC.Title	Title ⁺
Auteur(s) du corpus (affiliation): <i>(Par "l'auteur du corpus" on entend la ou les personne(s) qui ont créé le jeu de données et non la ou les personne(s) qui ont écrit les textes des documents contenus dans le corpus. Par exemple, si Jean-Paul Machin, chercheur au CNRS, a créé le corpus qui contient les textes de Dante Alighieri, on retient, comme auteur du corpus, J.-P.Machin et non Dante Alighieri.)</i>	DC.Creator	Authors ⁺
Contributeur(s) (affiliation, rôle): <i>(Il s'agit de toutes les personnes qui ont contribué, d'une façon ou d'une autre, à l'élaboration du corpus. Voir infra la liste des différents rôles possibles des contributeurs telle qu'elle est proposée par le dépôt Zenodo.)*</i>	DC.Contributeur	Contributors
Description (des données): <i>(Présenter ici le corpus, ces données, le programme de recherche qui a permis la construction de ce corpus, etc.)</i>	DC.Description	Description ⁺

Format de fichier (FAIR: I1): (Voir: Liste of file formats)	DC.Format	(Incorporer ces informations dans le champ "Description" lors de la saisie des métadonnées)
Logiciels spécifiques nécessaires pour la lecture des données: (nom, version, système d'exploitation) (FAIR: I1): <i>(Par exemple des logiciels de lexicométrie pour les données lexicostatistiques ou des logiciels de l'analyse de réseaux pour les données prosopographiques, etc. Les logiciels de traitement de texte (format doc, txt, csv, etc.), de bases de données (format sql), de visionnage des images (jpg, tiff, etc.), de lecture des fichiers XML (sauf des cas particuliers comme, par exemple, les logiciels lexicométriques TXM ou Philologic4 qui s'appuient sur un formatage XML particulier), etc. ne sont pas des logiciels spécifiques et ils ne doivent pas être mentionnés.)</i>		(Incorporer ces informations dans le champ "Description" lors de la saisie des métadonnées)
Si format ou logiciel particulier, expliquer clairement pourquoi ils ont été choisis (FAIR: I1):		(Incorporer ces informations dans le champ "Description" lors de la saisie des métadonnées)
Si les données ne sont pas explicites en l'état, renvoyer vers les explications/ documentation/vocabulaire pour expliciter les données (FAIR: I2):		(Incorporer ces informations dans le champ "Description" lors de la saisie des métadonnées)
Provenance et historique des données (FAIR R.1.2): <i>(Il s'agit d'indiquer l'historique et la provenance d'un jeu de données en tant qu'objet digital. Par exemple, les données peuvent être produites lors d'un autre projet et ensuite réutilisées, corrigées et modifiées par le présent projet. Le contenu de ce champ peut, par exemple, être le suivant: "Les données qui concernent les VIIIe et IXe siècles ont été extraites d'un jeu de donnée sous format sql produit lors du projet "X" et corrigées ensuite par Monsieur Untel. Les données qui concernent le XIIe siècle ont été extraites d'un jeu de donnée sous format xml produit lors du projet "Y" sans aucune correction supplémentaire. Toutes les autres données ont été saisies manuellement par Monsieur Unautre et Madame Uneautre sous format des documents JSON dans la base MongoDB". Pour plus de détail sur la notion de la provenance des données, voir par exemple: rd-alliance / FAIR-data-maturity-model-WG - Indicators for R1.2)</i>		(Incorporer ces informations dans le champ "Description" lors de la saisie des métadonnées)

Standards, conventions des données utilisées (FAIR: R1.3): <i>(Bien qu'un "standard" commun ne semble pas exister en sciences humaines et sociales, certains projets ou programmes peuvent s'appuyer sur des conventions similaires. Par exemple, l'organisation des données dans le même type de tableau, l'utilisation des mêmes principes de balisage, etc.)</i>		<i>(Incorporer ces informations dans le champ "Description" lors de la saisie des métadonnées)</i>
Version (des données): <i>(Il est conseillé d'utiliser la nomenclature du "versioning semantic". Pour plus de détails sur le "versioning semantic" voir Semantic Versioning)</i>		Version
Langue principale des données: <i>(Si, par exemple, les textes sont en latin, mais ils sont accompagnés par les analyses en français, on indique ici la langue principale des données qui est donc le latin.)</i>	DC.Language	Language
Mots-clés:	DC.Subject	Keywords
Droit d'accès et licence des données (FAIR: R1.1): <i>(Il existe différents types du droit d'accès et de licences. Voir infra la liste des droits d'accès proposée par le dépôt Zenodo. Le nombre de différentes licences est bien trop élevé pour les mentionner toutes.)**</i>	DC.Rights	Access right & License ⁺
Financement(s): <i>(Il s'agit de tous les financements qui ont permis l'élaboration de ce corpus.)</i>		Funding / Grants
Données liées (FAIR: I3): <i>(Mentionner, le cas échéant, le PID, l'intitulé des données liées et le type de lien. Par exemple: "ces textes (PID, intitulé du jeu de données avec les textes) est la transcription de ces photos de manuscrits (PID, intitulé du jeu de données avec les images) ".)</i>	DC.Relation	Related/alternate identifiers

⁺ Champ obligatoire dans le dépôt de données Zenodo.

* Liste des rôles des contributeurs (proposé par le dépôt Zenodo):

- Contact person
- Data collector
- Data curator
- Data manager
- Distributor
- Editor
- Hosting institution
- Other
- Producer
- Project leader
- Project manager

- *Project member*
- *Registration authority*
- *Related person*
- *Research group*
- *Researcher*
- *Rights holder*
- *Sponsor*
- *Supervisor*
- *Work package leader*

****** Les différents types des droits d'accès (proposé par le dépôt Zenodo):

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

Annexe 5

Évaluation complète (questionnaire détaillé) de la correspondance des données aux principes FAIR

Ce questionnaire peut-être utiliser lors de l'évaluation complète de la correspondance des données d'un corpus aux principes FAIR. À noter toutefois que ce questionnaire s'appuie sur l'hypothèse du versement des données dans un dépôt institutionnel de données quand la partie "métadonnées" est garantie avant tout par le dépôt. Dans cette optique, l'auteur des données garantit essentiellement les principes FAIR liés aux données. Cela explique le remplacement dans certaines questions du vocable "(méta)données" uniquement par le vocable "données" (voir *supra* "1.4 Qui FAIRise quoi?" et Annexes 1, 2, 3).

1. Données administratives
Nom du corpus
Adresse web du site où les données sont disponibles actuellement (sinon, adresse web du site du projet)
Responsable scientifique
Responsable technique (si différent)
2. Données techniques
Quelle est la nature des données (textuelle, visuelle, sonore, etc.)?
Quels sont les formats des données utilisés (txt, csv, xml, jpg, mysql, etc.)?
Quels sont les standards des données utilisés (e.g.: XML / DTD, modèle personnel, etc.)?
Quelle est la taille des fichiers contenant les données (en mb)?
Quelles sont les unités de données (par exemple actes, personnes, images) et combien d'unités contiennent les données? (Par exemple 4000 actes, 232 personnes, 59 images)?

Est-ce que des logiciels spécifiques sont nécessaires pour la lecture des données?

(Par exemple des logiciels de lexicométrie pour les données lexicostatistiques ou des logiciels de l'analyse de réseaux pour les données prosopographiques, etc.)

Les logiciels de traitement de texte (format doc, txt, csv, etc.), de bases de données (format sql), de visionnage des images (jpg, tiff, etc.), de lecture des fichiers XML (sauf des cas particuliers comme, par exemple, les logiciels lexicométriques TXM ou Philologic4 qui s'appuient sur un formatage XML particulier), etc. ne sont pas des logiciels spécifiques et ils ne doivent pas être mentionnés.)

Quelle est la plateforme d'affichage et de stockage des données (plateforme spécifique, CMS, site dédié, etc.)?

Par qui les données sont-elles hébergées?

(Par exemple un site institutionnel, un site personnel hébergé sur un serveur commercial, etc. Si les données ne sont pas en ligne, elles peuvent être stockées sur un ordinateur personnel)

3. Correspondance aux principes FAIR

F1: Les données possèdent-elles un identifiant persistant et unique au monde (PID)?

(Par exemple DOI, ARK, ISBN, etc.)

F2: Les données sont-elles décrites avec des métadonnées détaillées?

F3: Le PID des données (voir F1) est-il mentionné dans les métadonnées?

F4: Les données et les métadonnées y associées sont-elles enregistrées ou indexées dans une ressource consultable (par exemple, un dépôt de données)?

A1.1: Les données sont-elles accessibles via un protocole de communication ouvert, gratuit et largement utilisé?

(Il s'agit avant tout des protocoles informatiques libres et connus comme HTTP (=site web), FTP etc. Les autres "protocoles de communications" qui s'appuient sur des logiciels propriétaires et non universellement implantés (par exemple Skype, Microsoft Exchange Server, etc) ne répondent pas à ce critère).

A1.2: Les données sont-elles accessibles via un protocole de communication qui permet, si nécessaire, de mettre en place une procédure d'authentification et d'autorisation?

(FAIR data ne signifie pas OpenData. Le protocole via lequel les données et les métadonnées sont accessibles, doit avoir la possibilité, en cas de besoin, de restreindre l'accès aux données, par exemple uniquement aux utilisateurs possédant un login et un mot de passe.)

A2: Les métadonnées seront-elles accessibles même lorsque les données elles-mêmes ne seront plus disponibles?

(Normalement, ce critère est considéré comme rempli, si la condition F4 est satisfaite).

I1: Les données utilisent-elles un langage connu et largement utilisé?

(Par exemple XML, SQL, TXT, CSV, etc.)

I2: Est-ce que les données sont explicites et compréhensibles dans l'état? Si non, est-ce qu'elles sont accompagnées d'une documentation détaillée?

(Cela signifie que toutes les abréviations, tous les termes ambigus ou toutes les balises utilisées dans la description des données doivent être clairement expliqués ou avoir déjà une documentation existante).

I3: Dans le cas où les présentes données peuvent être reliées à d'autres données (par exemple, les données textuelles peuvent être reliées aux images des manuscrits), est-ce que les métadonnées contiennent les références à ces données reliées (par exemple le PID des données reliées, voir F1)? Si oui, est-ce que le type de lien entre les données est explicitement mentionné? (Par exemple, "ce texte est la transcription de ce manuscrit" ou "ce texte a été écrit par cette personne".)

R1.1: Est-ce que les données disposent d'une licence et est-ce que cette dernière est clairement mentionnée (par exemple dans les métadonnées)?

(Les données peuvent avoir tous types de licence, y compris les licences "non libres de droits").

R1.2: Est-ce que les données contiennent (par exemple dans les métadonnées) un historique détaillé de la provenance de ces données?

(Par exemple, les données peuvent être produites lors d'un autre projet et ensuite réutilisées, corrigées et modifiées par le présent projet).

R1.3: Est-ce que les données utilisent des standards de présentation, d'organisation ou d'archivage communément utilisés dans le domaine scientifique auquel ces données appartiennent?

(Bien qu'un "standard" commun ne semble pas exister en sciences humaines et sociales, certains projets ou programmes peuvent s'appuyer sur des conventions similaires. Par exemple, l'organisation des données dans le même type de tableau, l'utilisation des mêmes principes de balisage, etc. Si c'est le cas, ces conventions doivent être mentionnées, par exemple dans les métadonnées. Ne pas confondre avec I2. Le I2 vise à rendre les données compréhensibles / interopérables grâce à la documentation associée à ces données. Le R1.3 vise à faciliter la réutilisation des données grâce aux indications générales sur les standards et les conventions sur lesquels les données s'appuient.)

Annexe 6

Évaluation rapide (check-list) de la correspondance des données aux principes FAIR

Cette check-list peut-être utiliser lors de l'évaluation rapide de la correspondance des données d'un corpus aux principes FAIR.

Données

Les données possèdent un identifiant pérenne (par exemple DOI, ARK, etc.) [F1] ☐

Les données sont accompagnées par des métadonnées [F2] ☐

Les données sont indexées et accessibles (par exemple par l'intermédiaire d'un site internet) [F4, A1] ☐

Les données sont encodées dans un format et avec un langage informatique connus et répandus [I1] ☐

Les données sont accompagnées par une documentation
ou sont basées sur les standards déjà existants [I2, R1.3] ☐

Métadonnées

Les métadonnées sont indexées et accessibles (par exemple par l'intermédiaire d'un site internet) [F4, A1] ☐

Les métadonnées sont encodées sous des formats et selon des normes et des standards en vigueur
(par exemple RDF/XML, JSON-LD, etc.) [I1, I2, R1.3] ☐

Les métadonnées qui accompagnent les données contiennent, au minimum, les éléments suivants:

- identifiant pérenne des données auxquelles elles [métadonnées] sont reliées [F3] ☐
- description la plus complète possible des données (auteurs, intitulé, date de création, etc.) [F2] ☐
- licence d'utilisation des données [R1.1] ☐
- provenance des données [R1.2] ☐
- liens, le cas échéant, vers les autres données associées [I3] ☐

Annexe 7

FAIR par la pratique & Éléments-clés des principes FAIR

Ce rapide tour des éléments-clés des principes FAIR et de leur mise en pratique permet de livrer, quand cela est nécessaire, une présentation succincte de l'initiative FAIR.

voir page suivante

FAIR par la pratique

FAIRiser quoi?

Avant tout, il faut savoir quelles sont les données que l'on souhaite et que l'on peut FAIRiser. De fait, une *"unité de données"* à laquelle les principes FAIR sont appliqués est communément désignée comme un *"Objet Digital FAIR"* (*FAIR Digital Object*, FDO). C'est à ce FDO que les identifiants pérennes et les métadonnées sont associés. Mais, en pratique, comment choisir ce FDO?

Les principes FAIR indiquent seulement que le FDO doit être *"une entité significative"* pour la FAIRisation. Il appartient donc à chaque projet de faire le choix de *"l'unité"* de la FAIRisation. Dans les faits, **il est possible de FAIRiser soit un "jeu de données" (dataset) qui regroupe alors toutes les données d'un corpus, soit une "unité atomique" d'un corpus** (par exemple, une notice, un document, une image...). Il faut, toutefois, garder à l'esprit que la FAIRisation des unités atomiques, une par une, signifiera de déposer dans un dépôt des données (si cette approche est choisie) des centaines/milliers de fichiers, un par un. Et cela ne semble pas être raisonnable à l'échelle d'un corpus.

À noter également que des corpus constitués de données d'une nature différente (par exemple, données textuelles, visuelles, chiffrées, etc.) peuvent regrouper chaque type de données dans un jeu de données séparé. Autrement dit, **au sein d'un même corpus il peut exister plusieurs jeux de données où chaque "jeu" réunit des données analogues, par exemple les images, les textes, les notes, etc.**

FAIRiser comment?

Bien qu'il revienne à chaque projet de choisir les moyens par lesquels il souhaite atteindre les différents principes FAIR, **il est généralement possible d'envisager deux approches: la mise des données dans un dépôt de données institutionnel ou l'élaboration d'une infrastructure personnelle de gestion des données.** Si, pour une raison ou une autre, vous voulez garder la main sur la totalité du processus et si vous avez la possibilité de faire accomplir les tâches techniques lourdes par du personnel compétent vous pouvez opter pour l'infrastructure personnelle, dans tous les autres cas, il vaut mieux se tourner vers un dépôt des données.

Pour plus de détails sur les différentes approches de la FAIRisation, voir: [Écosystème FAIR](#)

Exemple comment un dépôt peut remplir les critères FAIR: [Zenodo: FAIR Principles](#)

Exemple de modèle des données pour un dépôt institutionnel: [Dataset upload example](#)

Exemple d'un dépôt réalisé: [CartuLR. Répertoire des cartulaires médiévaux et modernes](#)

Éléments-clés des principes FAIR

Les recommandations FAIR sont composées de 15 principes (ou 27 dans la version étendue, voir: [Principes FAIR étendus](#)) qui doivent être satisfaits afin que les données soient considérées comme "FAIRisées". Il est cependant possible de résumer l'ensemble de ces principes en quelques éléments-clés.

Identifiant pérenne

Principes FAIR concernés:

F1, F3, A1

Points à retenir:

☞ Il existe différents systèmes d'identifiants pérennes (*Persistent Identifier, PID*). Parmi les plus connus, on peut citer DOI ([Digital Object Identifier](#)), ARK ([Archival Resource Key](#)) ou bien encore VIAF ([Virtual International Authority File](#)). Par exemple, un DOI se présente comme "10.1000/182", un ARK comme "ark:/53355/cl010066723" et un VIAF comme "106965171". Il est possible de choisir un type d'identifiant qui convient le mieux à vos données. À noter toutefois qu'un DOI reste l'identifiant générique le plus utilisé et le mieux connu et qu'il peut être aisément associé à une plus grande variété de données numériques.

☞ Une URL (*Uniform Resource Locator*, communément appelée "adresse web") peut, à certains égards, être considérée comme un PID, mais son utilisation doit être déconseillée à cause de son caractère instable. En pratique: Si vos données sont, par exemple, stockées et affichées par le biais de votre propre site internet, l'adresse web (URL) qui permet d'accéder à la page où vos données sont affichées peut, à certains égards, être considérée comme un identifiant pérenne pour les données en question. Or, cette approche est contre-indiquée car la nature même d'un site internet ne permet pas de garantir la persistance d'une adresse web à moyen et, surtout, à long terme.

☞ Toutes les données que vous souhaitez FAIRiser doivent posséder un identifiant pérenne. Si vous optez pour la FAIRisation de votre corpus en tant que seul jeu de données c'est ce jeu de données qui doit avoir un PID. Si vous choisissez de FAIRiser les notices de votre corpus une par une alors c'est chaque notice qui doit détenir son propre PID.

Comment procéder?

dépôt de données: Si vous entreposez les données dans un dépôt de données, le dépôt se charge automatiquement de la création et de l'attribution d'un PID pour vos données.

infrastructure personnelle: Si c'est l'infrastructure personnelle qui a été choisie pour la FAIRisation, il vous revient de créer et d'attribuer par vous-même un PID pour vos données. (Pour la création manuelle d'un identifiant pérenne, vous pouvez vous appuyer, par exemple, sur les services de [OPIDoR-CNRS](#)).

Métadonnées

Principes FAIR concernés:

F1, F2, F3, F4, A1, I1, I2

Points à retenir:

- ☞ Toutes les données que vous souhaitez FAIRiser doivent être accompagnées par des métadonnées.
- ☞ Dans la plupart des cas, les métadonnées sont stockées dans un fichier spécifique. Ce fichier doit être encodé sous un format et selon les standards recommandés pour la représentation des métadonnées. Parmi ces formats et standards, on retient avant tout les recommandations RDF et les formats RDF/XML et JSON-LD.
- ☞ Chaque fichier de métadonnées doit contenir l'identifiant pérenne (PID) des données auxquelles elles [métadonnées] sont associées.
- ☞ Les métadonnées doivent contenir des informations suffisamment détaillées (auteurs, intitulé, date de création, description, mots-clés, etc.) sur les données qu'elles accompagnent. Il est possible, par exemple, de s'appuyer sur les éléments du vocabulaire *Dublin Core* pour la construction des métadonnées.

Comment procéder?

dépôt de données: Si vous entreposez les données dans un dépôt de données, le dépôt se charge automatiquement de la création et de la conservation des fichiers des métadonnées avec un format et selon les standards en vigueur. Il ne vous reste qu'à remplir les champs (auteurs de données, contenu du dépôt, etc.) fournis par le dépôt lors de l'ajout de vos données.

infrastructure personnelle: Si c'est l'infrastructure personnelle qui a été choisie pour la FAIRisation, il vous revient de vous assurer que vos métadonnées contiennent tous les éléments nécessaires, qu'elles sont conservées par vos soins sous un format et selon les standards recommandés et qu'elles [métadonnées] sont reliées (par le biais d'un PID) à vos données.

Stockage et affichage des données et des métadonnées

Principes FAIR concernés:

A1, A2

Points à retenir:

☞ Les données et les métadonnées doivent être accessibles par un moyen répandu et connu du plus grand nombre. En pratique, on s'attend, dans la majorité des cas, que les données et les métadonnées soient accessibles sur Internet par l'intermédiaire d'un site web.

Comment procéder?

dépôt de données: Si vous entreposez les données dans un dépôt de données, le dépôt met automatiquement à votre disposition un espace de stockage (un serveur) et un espace d'affichage (une page web) qui contient toutes les informations nécessaires pour accéder à vos données et métadonnées.

infrastructure personnelle: Si c'est l'infrastructure personnelle qui a été choisie pour la FAIRisation, il vous revient de créer et de maintenir un espace de stockage et d'affichage pour vos données et métadonnées.

Format et langage des données

Principes FAIR concernés:

I1

Points à retenir:

☞ Les données doivent être encodées sous un format et avec un langage connu et largement répandu. En pratique, c'est le critère le plus facile à satisfaire; il suffit d'éviter de représenter les données sous un format qui n'a été utilisé que par trois personnes dans le monde au milieu des années 80 du siècle passé...

Comment procéder?

auteur des données: Quelle que soit l'approche que vous choisissiez pour la FAIRisation de vos données (par l'intermédiaire d'un dépôt institutionnel ou par le biais d'une infrastructure personnelle), c'est à l'auteur de données qu'il appartient de s'assurer que ces données soient encodées sous un format et avec un langage appropriés.

Documentation des données

Principes FAIR concernés:

I2, R1.3

Points à retenir:

- ☞ Toutes les données doivent être explicites et compréhensibles en l'état. En pratique, cela signifie que le modèle et la structure des données, les abréviations, les vocabulaires et les conventions utilisés, les éventuels écueils et les difficultés du traitement des données, tous ces éléments doivent être décrits et annotés. En résumé, cette documentation doit être suffisamment complète pour permettre à une personne tierce de comprendre et d'analyser les données présentes.
- ☞ Si vous vous appuyez sur les modèles des données déjà décrits et documentés, par exemple les recommandations TEI P5 pour le langage XML, il suffit, dans la plupart des cas, de renvoyer vers cette documentation existante.

Comment procéder?

auteur des données: Quelle que soit l'approche que vous choisissiez pour la FAIRisation de vos données (par l'intermédiaire d'un dépôt institutionnel ou par le biais d'une infrastructure personnelle), c'est à l'auteur de données qu'il appartient de mettre en place une documentation appropriée.

Licence d'utilisation

Principes FAIR concernés:

R1.1

Points à retenir:

- ☞ Les métadonnées doivent contenir de façon claire et explicite la licence d'utilisation associée aux données qu'elles [métadonnées] accompagnent.

Comment procéder?

dépôt de données: Si vous entreposez les données dans un dépôt de données, vous pouvez choisir la licence lors de la mise en ligne; ce choix sera retenu par le dépôt et ajouté dans les métadonnées générées automatiquement.

infrastructure personnelle: Si c'est l'infrastructure personnelle qui a été choisie pour la FAIRisation, il vous revient de vous assurer que vos métadonnées contiennent, de façon claire et explicite, la licence d'utilisation appropriée.

Provenance des données

Principes FAIR concernés:

R1.2

Points à retenir:

☞ Les métadonnées doivent contenir les informations sur la provenance des données. En pratique, il s'agit, le plus souvent, d'indiquer la provenance de données en tant qu'objet digital. Par exemple, les données peuvent être produites lors d'un autre projet et ensuite réutilisées, corrigées et modifiées par le présent projet. Le contenu de ce champ peut, par exemple, être le suivant: *"Les données qui concernent les VIII^e et IX^e siècles ont été extraites d'un jeu de donnée sous format SQL produit lors du projet "X" et corrigées ensuite par Monsieur Untel. Les données qui concernent le XII^e siècle ont été extraites d'un jeu de donnée sous format XML produit lors du projet "Y" sans aucune correction supplémentaire. Toutes les autres données ont été saisies manuellement par Monsieur Unautre et Madame Uneautre sous format des documents JSON dans la base MongoDB"*. Pour plus de détail sur la notion de la provenance des données, voir par exemple: [*RD-Alliance/Issues: Indicators for R1.2: \(meta\)data are associated with detailed provenance*](#).

Comment procéder?

dépôt de données: Si vous entreposez les données dans un dépôt de données, vous pouvez soit inclure les informations sur la provenance des données dans le champ "Description" lors de la mise en ligne, soit accompagner vos données avec un document explicatif (par exemple un fichier "Readme") qui contiendra ces informations.

infrastructure personnelle: Si c'est l'infrastructure personnelle qui a été choisie pour la FAIRisation, vous pouvez soit inclure les informations sur la provenance des données dans vos métadonnées, soit accompagner vos données avec un document explicatif (par exemple un fichier "Readme") qui contiendra ces informations.

Annexe 8

Exemple de FAIRisation par l'intermédiaire du versement de données dans un dépôt

Rappel: L'exemple de la FAIRisation qui suit se base entièrement sur le programme de la FAIRisation proposé plus haut (voir *supra* "2. Étapes de la FAIRisation"). Chaque étape de la FAIRisation a été appliquée au corpus "[Comptabilités Principautés S / E](#)". Le dépôt retenu pour le versement des données est Zenodo (administré par le CERN et soutenu par la Commission européenne et les programmes Horizon 2020 et OpenAIRE). (Sur le choix d'un dépôt voir *supra* Annexe 3)

À noter cependant qu'il s'agit de la FAIRisation "allégée". De fait, le corpus "Comptabilités Principautés S / E" contient aussi bien les textes que les images; la FAIRisation qui suit ne s'applique qu'aux textes. On retient également que dans l'hypothèse du versement des données dans un dépôt institutionnel, la partie "métadonnées" des principes FAIR est garantie avant tout par le dépôt. Dans cette optique, l'auteur des données garantit essentiellement les principes FAIR liés aux données (voir *supra* "1.4 Qui FAIRise quoi?" et les Annexes 1 et 2).

Le dépôt Zenodo possède le mécanisme de la gestion des versions (le versioning), ce qui permet, par la suite, de corriger, d'ajouter et de modifier les données au fur et à mesure. À noter toutefois que la mise à jour du dépôt passe ne pas par la correction des fichiers déjà déposés, mais par la création d'une nouvelle version avec son propre identifiant (DOI). Le jeu de données déposé possédera également un identifiant (DOI) "global" qui regroupe toutes les versions. Sur le versioning dans le dépôt Zenodo voir: [Zenodo - DOI versioning](#).

Pour le dépôt qui résulte de cette FAIRisation voir: [Ressources comptables en Dauphiné, Provence, Savoie et Venaissin \(XIIIe - XVe siècle\)](#)

Pour d'autres exemples des dépôts des corpus médiévaux, voir:

- He, Sheng, Schomaker, Lambert, Samara, Petros, & Burgers, Jan. (2016). MPS Data set with images of medieval charters for handwriting-style based dating of manuscripts (Version v1.0) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.1194357>

- Koho, Mikko, Tuominen, Jouni, Lewis, David, Ikkala, Esko, Heller, Benjamin, Thomson, Emma, ... Fraas, Mitch. (2021). Mapping Manuscript Migrations Knowledge Graph (Version 2.2.0) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.4440464>
- Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, & Mirko Tavoni. (2020). Two Datasets for the Computational Authorship Analysis of Medieval Latin Texts (Version 2.00) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.4298503>

voir page suivante

Présentation des données

(voir fiche détaillée du corpus "[Comptabilités Principautés S / E](#)")

Corpus: Comptabilités Principautés S / E

Format des données: XML (texte)

Standard des données: XML TEI P5

Unité de données: 25 comptes (=25 fichiers XML)

Affichage et stockage des données: les données sont affichées sur un site web personnel, mais sont stockées et interrogées via la base eXist-db.

FAIRisation des données

(voir supra le programme complet et toutes les étapes de la FAIRisation "2. Étapes de la FAIRisation")

1. Vérification des données

1.1. S'assurer que les langages et les formats/extensions utilisés pour l'encodage des données sont connus, largement utilisés et, de préférence, libres.

XML est un format connu et largement utilisé.

FAIR: I1 - rempli.

1.2. S'assurer que les données sont explicites et compréhensibles en l'état. Cela signifie que toutes les abréviations, tous les termes ambigus ou toutes les balises utilisées dans la description des données doivent être clairement expliqués ou avoir déjà une documentation existante.

XML TEI P5 est un standard connu qui possède une documentation.

FAIR: I2 - rempli.

1.3. Créer un jeu de données du corpus. (Le plus souvent c'est un seul fichier de l'archive numérique, par exemple sous format zip).

On crée un fichier de l'archive numérique zip de tous les fichiers XML.

2. Préparation des métadonnées

2.1. Préparation des métadonnées complètes.

2.1.1. S'assurer que les métadonnées contiennent le plus de détails possible. *(Se référer au modèle des métadonnées détaillées, voir Annexe 4)*

Il n'existe aucune métadonnée pour le jeu de données du corpus. On écrit (pour l'instant dans n'importe quel fichier texte) les métadonnées en se basant sur le modèle de l'Annexe 4. Ces métadonnées seront par la suite utilisées lors de la mise des données dans un dépôt.

FAIR: F2 - en attente.

2.1.2. S'assurer que les données disposent une licence et qu'elle est clairement mentionnée dans les métadonnées. *(Les données peuvent avoir tous types de licence, y compris les licences "non libres de droits")*.

Si cela n'a pas été encore fait, on ajoute les informations sur la licence dans le brouillon des métadonnées créées précédemment.

Licence des données (corpus "Comptabilités Principautés S / E"): Open Access (Creative Commons Attribution 4.0 International).

FAIR: R1.1 - en attente.

2.1.3. S'assurer que les métadonnées contiennent un historique détaillé de la provenance des données. *(Par exemple, les données peuvent être produites durant un autre projet et ensuite réutilisées, corrigées et modifiées par le présent projet).*

Si cela n'a pas été encore fait, on ajoute les informations sur la provenance des données dans le brouillon des métadonnées créées précédemment.

FAIR: R1.2 - en attente.

2.1.4. S'assurer que les données utilisent des standards (s'il en existe) de présentation, d'organisation ou d'archivage communément utilisés dans le domaine scientifique auquel ces données appartiennent. *(Ces informations seront par la suite utilisées dans les métadonnées).*

Si cela n'a pas été encore fait, on ajoute les informations sur les standards dans le brouillon des métadonnées créées précédemment.

Standards, conventions des données utilisées (corpus "Comptabilités Principautés S / E"): XML TEI P5.

FAIR: R1.3 - en attente.

2.2. S'assurer que dans les cas où les présentes données peuvent être reliées à d'autres données (par exemple, les données textuelles peuvent être reliées aux images des manuscrits), les métadonnées contiennent des références à ces données reliées (par exemple le PID des données reliées, voir F1). Le type de lien entre les données doit être explicitement mentionné. (Par exemple, "ce texte est la transcription de ce manuscrit" ou "ce texte a été écrit par cette personne").

Si cela n'a pas été encore fait, on ajoute les informations sur les données liées dans le brouillon des métadonnées créées précédemment.

FAIR: I3. - en attente.

3. Mise des données dans un dépôt

3.1. Enregistrement des données et des métadonnées.

3.1.1. Choisir une ressource consultable (par exemple, un dépôt de données) qui répond à tous les critères FAIR applicables aux métadonnées (F1, F3, F4, A1.1, A1.2, A2, I1, I2, I3, R1.1, R1.2, R1.3). La plupart des grands dépôts de données affichent de façon claire leur conformité aux critères FAIR. *(Pour plus de détails sur le choix de dépôt et sur les critères FAIR auxquels il doit répondre voir infra Annexe 3)*

On choisit le dépôt Zenodo. Ce dépôt de données est une infrastructure officielle hébergée par le CERN et financée par l'European Commission (OpenAIRE & Horizon 2020). Ce dépôt répond à tous les critères FAIR, voir [Zenodo-FAIR Principles](#)

3.1.2. Enregistrer les données et les métadonnées y associées dans une ressource consultable choisie précédemment.

On enregistre les données et les métadonnées dans le dépôt Zenodo.

FAIR: F4 - rempli.

FAIR: F2 - rempli.

FAIR: R1.1 - rempli.

FAIR: R1.2 - rempli.

FAIR: R1.3 - rempli.

FAIR: I3. - rempli.

voir les captures d'écran pages suivantes

captures d'écran

Ajouter le jeu de données

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

Drag and drop files here

— or —

Choose files

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

If you're experiencing issues with uploading larger files, read our [FAQ section](#) on file upload issues.

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
Ressources comptables en Dauphiné, Provence, Savo ...	1.3 MB		

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)

If you're experiencing issues with uploading larger files, read our [FAQ section](#) on file upload issues.

Remplir le formulaire des métadonnées

Upload type required ▼

Publication ○	Poster ○	Presentation ○	Dataset ●	Image ○	Video/Audio ○	Software ○	Lesson ○	Physical object ○	Other ○
------------------	-------------	-------------------	--	------------	------------------	---------------	-------------	----------------------	------------

Basic information required ▼

Digital Object Identifier

Publication date *

Title *

Authors *

Description *

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.

Reserve DOI

2021-06-11

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

Ressources comptables en Dauphiné, Provence, Savoie et Venaissin (XIIIe - XVe siècle)

Required.

Jamme, Armand

UMR 5648, CNRS

ORCID (e.g.: 0000-0002-1825-0097) ↕ ✕

+ Add another author

Ce corpus, né d'un programme financé par l'Agence nationale de la Recherche, intitulé «Genèse médiévale d'une méthode

Version

Language

Keywords

Additional notes

1.0.0

Optional. Mostly relevant for software and dataset uploads. Any string will be accepted, but semantically-versioned tag is recommended. See [semver.org](#) for more information on semantic versioning.

Latin

Optional. Primary language of the record. Start by typing the language's common name in English, or its ISO 639 code (two or three-letter code). See [ISO 639 language codes list](#) for more information.

histoire médiévale ↕ ✕

comptabilité médiévale ↕ ✕

comptes des châtelains ↕ ✕

+ Add another keyword

Optional.

License
required

Access right *

☒ Open Access

☐ Embargoed Access

☐ Restricted Access

☐ Closed Access

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

License *

Creative Commons Attribution 4.0 International

Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the *Other* licenses available (*Other (Open)*, *Other (Attribution)*, etc.). The supported licenses in the list are harvested from opendefinition.org and spdx.org. If you think that a license is missing from the list, please [contact us](#).

Funding
recommended

Zenodo is integrated into reporting lines for research funded by the European Commission via [OpenAIRE](#). Specify grants which have funded your research, and we will let your funding agency know!

Grants

Agence Nationale de la Recherche (FR)

GEMMA ANR-10-BLAN-2011 Medieval Genesis of an Administrative Metho

Optional. OpenAIRE-supported projects only. For other 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](#)

Contributors
optional

Contributors

Cereia, Daniela	Archives d'État de Turin	ORCID (e.g.: 0000-0002-1)	Project member	Optional.
Lemonde, Anne	Université de Grenoble-II	ORCID (e.g.: 0000-0002-1)	Project member	Optional.
Vallière, Laurent	UMR 5648, CNRS	ORCID (e.g.: 0000-0002-1)	Project member	Optional.
Hayez, Anne-Marie	rtementales de Vaucluse	ORCID (e.g.: 0000-0002-1)	Project member	Optional.
Burghart, Marjorie	UMR 5648, CNRS	ORCID (e.g.: 0000-0002-1)	Project member	Optional.
Grunin, Andrey	UMR 5648, CNRS	ORCID (e.g.: 0000-0002-1)	Project member	Optional.

Uploader le fichier du corpus

Start upload

Publier le dépôt en ligne

Publish

Le dépôt est en ligne: <https://zenodo.org/record/4919334>

3.2. Vérification de l'accessibilité et de l'exactitude des (méta)données.

3.2.1. Vérifier que les données possèdent un identifiant persistant et unique (PID) (fourni par la ressource où les données sont enregistrées).

Le jeu de données possède l'identifiant DOI ("global"):

<http://doi.org/10.5281/zenodo.4919334>

FAIR: F1 - rempli.

3.2.2. Vérifier que le PID des données est mentionné dans les métadonnées.

Sur la page du dépôt on choisit Export, par exemple JSON:



Dans le texte qui s'affiche, on trouve l'identifiant:

```
"metadata": {  
  "access_right_category": "success",  
  "doi": "10.5281/zenodo.4919335",  
  "version": "1.0.0",
```

FAIR: F3 - rempli.

3.2.3. Vérifier que les données sont accessibles via un protocole de communication ouvert, gratuit et largement utilisé. Si les données sont enregistrées dans une ressource qui est accessible via un site internet, ce critère est automatiquement rempli.

Le jeu de données est accessible via le dépôt Zenodo (voir [Zenodo-FAIR Principles](#)).

FAIR: A1.1 - rempli.

3.2.4. Vérifier que les données sont accessibles via un protocole de communication qui permet, si nécessaire, de mettre en place une procédure d'authentification et d'autorisation.

Si nécessaire, il est possible de restreindre l'accès aux données mises en ligne dans le dépôt Zenodo (voir [Zenodo-FAIR Principles](#)).

FAIR: A1.2 - rempli.

3.2.5. Vérifier que les métadonnées seront accessibles même lorsque les données elles-mêmes ne seront plus disponibles. *(De façon ordinaire ce critère est considéré comme rempli, si la condition F4 est satisfaite).*

Le dépôt Zenodo garantit l'accessibilité des métadonnées durant toute l'existence du dépôt (voir [Zenodo-FAIR Principles](#)).

FAIR: A2 - rempli.