Vademecum pour la réutilisabilité des données

Groupe de travail *Réutilisatibilité*, Consortium Cahier
Janvier 2022

Vademecum rédigé par Julie Aucagne, Marguerite Bordry, Camille Desiles, Francine Filoche, Anne Garcia-Fernandez, Elisabeth Greslou, Camille Koskas, Gwenaëlle Patat, Richard Walter et Pierre Willaime.

Table des matières

1 Voir https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf

Les liens dans le texte sont signifiés en couleurs. Ils sont aussi rappelés en note. Certains termes techniques ont une entrée dans le glossaire en annexe C : la première occurrence d'un terme défini dans celui-ci est également en couleur mais pour ne pas saturer le texte, nous n'avons pas indiqué les autres occurrences.

1 Introduction

Ce document est issu des réflexions du groupe de travail « (Ré)utilisabilité » du consortium Cahier[2], consortium de projets d'édition de corpus d'auteurs, qui s'est réuni plusieurs fois en 2021. Son objectif est de proposer des recommandations pratiques concernant la réutilisabilité des données textuelles dans le cadre des projets de recherche.

Ce travail s'inscrit, pour le contexte français, dans les démarches d'incitation à l'ouverture des données de la recherche[3] initié par la Loi pour une République numérique[4] de 2016 et encouragé par le mouvement de la Science Ouverte, dont le premier plan[5] a été publié en 2018. Le deuxième plan national pour la Science ouverte[6], qui couvre la période 2021-2024, insiste encore davantage sur la notion de réutilisation en incitant à « reconnaître et amplifier la réutilisation des données de la recherche ».

Si les trois premières lettres des principes FAIR[7] sont de plus en plus intégrées dans les projets numériques, le R – réutilisabilité/réutilisation –, qui découle de ces trois premières lettres, est encore peu mis en pratique réellement. Pourtant, l'une des premières étapes de tout projet de recherche est de vérifier les données disponibles pour la constitution de ses sources. On se heurte alors souvent à la collecte de données hétérogènes qu'il faut aligner et harmoniser pour pouvoir concrètement les exploiter. Suite à ce constat, le groupe a décidé de rédiger des préconisations afin que le « R » de FAIR soit davantage pris en compte et mis en œuvre dès le départ de tout projet numérique.

La notion de réutilisation est vaste. Les données peuvent être réutilisées dans des optiques pédagogiques, patrimoniales, par d'autres disciplines, mais aussi enrichies dans le cadre de recherches complémentaires. La réutilisation peut être envisagée dans ses aspects méthodologiques : on peut réutiliser aussi des choix éditoriaux, des développements, des processus…

Les données de la recherche sont ici entendues à la fois comme les sources mais aussi comme les données transformées, les métadonnées, la documentation, les scripts et les codes liés[8] au projet. La réutilisation peut donc porter aussi bien sur les données elles-mêmes (collectées ou produites), que sur les processus et outils utilisés sur celles-ci tout au long du projet.

2 Voir https://cahier.hypotheses.org
3 Voir https://doranum.fr/glossaire-donnees-recherche
4 Voir https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829
5 Voir https://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html
6 Voir https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte
7 FAIR : Faciles à (re)trouver, Accessibles, Interopérables, Réutilisables.
8 Voir https://www.legifrance.gouv.fr/circulaire/id/45162

Après avoir dressé les constats sur les pratiques et les verrous autour de la réutilisation des données, notre *vademecum* propose des modalités de mise en œuvre en amont puis tout au long du projet.

Le présent document comporte des annexes dans lesquelles figurent un glossaire, une bibliographie, un ensemble de « trucs & astuces » ainsi qu'une *check-list* avec des critères à remplir pour favoriser la réutilisabilité des données.

Le présent document n'aborde qu'en marge la question de l'archivage des données qui soulève des questions spécifiques[9]. Nous nous limitons ainsi à considérer la réutilisation à court ou moyen terme.

9 Le Cines est l'un des principaux points d'entrée pour cette thématique : https://www.cines.fr/archivage

2 Pratiques et verrous de la réutilisation des données

2.1 La visibilité des données

Notre premier constat est qu'il semble plus aisé de trouver des données sur les sites des projets qu'à partir d'entrepôts ou de portails. En recherchant des corpus sur des thèmes spécifiques, on trouve plusieurs sites qui proposent des liens de téléchargement des données sous différents formats comme le format TEI[10] ou des formats bruts (avec le texte seul), des liens vers les codes sources, de la documentation… Il est fort utile de pouvoir accéder à toutes ces informations de façon immédiate. Mais les pratiques éditoriales varient de site en site. Ainsi, certains ne proposent pas de version brute des données mais une version pour tel ou tel logiciel.

Notre second constat est que les données issues de sites web sont dispersées et, ainsi, souvent peu visibles. En effet, il n'existe pas de catalogue, d'annuaire ou de cartographie qui recenseraient les jeux de données disponibles et produits dans le cadre d'un projet de recherche publique. Quand une initiative de ce genre est lancée, le processus rencontre de nombreux obstacles : il est nécessaire de trouver la bonne échelle et la bonne granularité pour recenser ces projets. Quel niveau de précision choisir dans la description de ces données (conditions d'accès, états et formats des données) ? Une trop grande précision implique des ressources trop importantes en temps et en énergie, mais une description trop vague ne constituerait pas un véritable apport pour la recherche. Les institutions ou les consortiums recensent leurs propres projets (CAHIER[11], OBTIC[12]), mais cela reste parcellaire. Différents catalogues de bibliothèques permettent de rassembler des sources, comme le SUDOC[13], mais ne donnent pas directement accès aux ressources et ne pointent pas vers les projets en cours de développement.

D'autre part, les entrepôts de données ne sont pas connus de tous les acteurs et, en conséquence, ne sont pas assez utilisés. Quant aux projets qui déposent leurs données sur un entrepôt, il est rare qu'ils rendent visible le lien entre ces données et leur contexte scientifique. Se pose donc le problème de la présence d'une documentation sur le projet, aussi bien sur la construction du projet scientifique que sur les données elles-mêmes. Comment ont-elles été fabriquées, quels sont les biais éditoriaux des auteurs ? On note aussi un manque de disponibilité des données brutes : on trouve les résultats finaux des projets (des éditions numériques, des visualisations, des sites internet, etc.), mais sans accès aux

10 Text Encoding Initiative (https://tei-c.org).
11 Voir https://cahier.hypotheses.org
12 Voir https://obtic.sorbonne-universite.fr. OBTIC a pris la suite d'OBVIL (http://obvil.sorbonne-universite.site).
13 Voir http://www.sudoc.abes.fr/cbs

sources au format TEXT ou TEI. Quand on les trouve, leur encodage est souvent trop lourd ou trop spécifique pour permettre une réutilisation.

On constate enfin l'absence de référencement des outils nécessaires à la gestion numérique des données ; ces outils ne donnent pas souvent lieu à des publications expliquant leurs avantages et inconvénients. Le choix des outils, bien que crucial, devient un moment compliqué du projet numérique, en particulier pour assurer la réutilisabilité des données.

2.2 Les obstacles juridiques

L'un des principaux verrous pour la réutilisation des données concerne les aspects juridiques. En France, le cadre légal prône l'ouverture des données au maximum (cf. loi sur la république numérique[14] et loi Valter[15]). En même temps, d'autres principes s'appliquent : les droits d'auteur et les droits voisins, la question de la propriété intellectuelle ou des données personnelles (droits à l'image, respect de la vie privée, etc.). C'est surtout flagrant pour les corpus récents (XX^{e}-XXI^{e} siècle). Il en découle des restrictions pour la réutilisabilité des données et leur citabilité[16]. Il existe une zone de flou, qui peut déstabiliser les porteurs de projets et les utilisateurs : par exemple, du point de vue des droits, certaines données sont inaccessibles, alors que leurs métadonnées sont, elles, accessibles.

Dans quelle mesure les métadonnées sont-elles concernées par le droit d'auteur ? Sans doute faudrait-il distinguer les métadonnées purement descriptives de celles qui sont le fruit d'un travail d'interprétation[17]. Mais si ce travail d'interprétation n'est pas signé, il est difficile concrètement de le protéger. Une des premières actions d'un projet doit être de chercher les indications de signatures et de protection juridiques (licence, source, propriété, etc.) des données. Si ces indications ne permettent pas la réutilisation des données, il est toujours possible de contacter les propriétaires des droits.

En négociant avec une institution ou un ayant droit, les porteurs de projet ne pensent pas forcément à évoquer avec les différents acteurs de ce cadre légal la question de la possible réutilisation des données et de leur dépôt sur un entrepôt.

Dans nos pratiques de recherche de corpus, nous découvrons souvent plusieurs sites présentant des données. La plupart indiquent une licence, mais qui n'est pas toujours ouverte. La raison n'est souvent pas précisée. Nous pouvons même rencontrer des sites proposant

14 Voir https://www.economie.gouv.fr/republique-numerique
15 Voir https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000031701525
16 Marie-Luce Demonet, « La Confiscation des données issues de l'humanisme numérique » *in* Véronique Ginouvès ; Isabelle Gras, *La Diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques*, Presses universitaires de Provence, 2018, Digitales, 9791032001790 ⟨hal-02068085⟩.
17 Le droit d'auteur s'applique à tout travail témoignant de l'originalité d'une pensée. Voir le tutoriel de Mate-SHS avec une intervention de Lionel Maurel, « À qui appartiennent les données ? » : https://mate-shs.cnrs.fr/actions/tutomate/tuto25-propriete-donnees-lionel-maurel

deux informations contradictoires : une licence *Creative Commons* permettant une réutilisation classique mais avec une métadonnée « tous droits réservés ».

2.3 Les facteurs humains

Les obstacles humains à la mise en œuvre de la réutilisabilité sont eux aussi divers : impression d'être dépossédé de son travail, préjugés envers la réutilisabilité considérée comme signe d'une mauvaise qualité des données car « en libre-service », inquiétude face à l' « accaparement » (*scraping*), compétition entre chercheurs, laboratoires ou institutions, privés comme publics. Ces facteurs peuvent mener à une autocensure face à la nécessité de donner les clés de réutilisation de ses données.

Il faut prendre en compte la multiplicité des acteurs se succédant trop souvent sur un projet (multiplication des CDD, appels à projets limités dans le temps, appel à des prestataires), ce qui peut amener à des problèmes de communication, d'autant plus que tous les acteurs d'un même projet ne partagent pas toujours le même vocabulaire. Le long terme peut poser problème pour les projets : on constate souvent l'absence d'un référent humain portant la mémoire du projet. Qui plus est, les données susceptibles d'être mises à disposition sont en évolution permanente : comment réutiliser des données en mutation constante ? Enfin, l'utilisateur peut être déstabilisé si le projet ne communique pas de manière transparente et régulière sur l'état des données disponibles.

Les choix scientifiques ne sont pas toujours explicités, alors qu'ils représentent un élément essentiel de chaque projet, potentiellement nécessaire pour la réutilisabilité. On trouve trop peu de Plans de gestion des données (PGD), de *data papers*, d'explications des choix techniques dans les articles publiés par les chercheurs, en partie parce que ce travail demeure trop peu valorisé. Ces choix doivent aussi être présents sur les sites diffusant les données, à proximité de celles-ci.

Les porteurs de projet peuvent aussi être échaudés par l'absence de retours de la part des utilisateurs de leurs données, alors que ces retours sont un facteur d'amélioration de la qualité et de la visibilité des données produites et *de facto* de la recherche les ayant produites.

2.4 Les verrous techniques

D'un point de vue technique, il existe un triple verrou aussi bien pour rendre les données réutilisables que pour les réutiliser. Premièrement les données sont rarement accessibles sous une forme brute qui les rendrait facilement exploitables. Deuxièmement, les données éditorialisées sont souvent disponibles sans documentation et sans contexte, alors qu'un ac-

cès à la documentation ou au schéma permettrait de comprendre les choix scientifiques et techniques. Troisièmement, la question des formats est essentielle : les formats propriétaires s'imposent trop souvent dans la pratique par rapport aux formats libres.

Les entrepôts de données permettent de récupérer de la donnée, mais si on ne récupère pas les métadonnées de celle-ci, il en résulte une perte d'information et de contextualisation. Le protocole OAI-PMH, qui gère un entrepôt avec uniquement les métadonnées, est très utilisé pour assurer la réutilisabilité mais il ne faut pas perdre le lien entre les métadonnées et la donnée. Le protocole fonctionne par un jeu de question/réponse avec l'entrepôt ; dans la réponse le lien avec la donnée source est toujours indiqué. Il faut alors trouver le moyen de consulter ou de récupérer la donnée par ce lien. Par là même, nous avons déjà les métadonnées avant même la donnée. L'OAI-PMH est ainsi une bonne solution pour augmenter la visibilité et la réutilisation de ses données via les métadonnées.

Par ailleurs, une des barrières au dépôt d'une donnée dans un entrepôt est le passage d'un type de métadonnées à un autre (celles présentes dans le corpus et celles formatées de l'entrepôt choisi), et donc les risques potentiels d'atténuation de leur finesse, voire de la perte de leur pertinence. Ainsi, si l'on dispose d'un fichier TEI avec un entête (teiHeader) très riche contenant une grande quantité de métadonnées, il n'est pas simple de savoir comment conserver cette richesse au sein d'un entrepôt qui ne reposerait uniquement sur le standard de métadonnées Dublin Core. Il faut alors prévoir un stockage des autres métadonnées au sein même de la donnée.

Enfin, quand les données sont disponibles, l'utilisateur peut rencontrer des difficultés pour évaluer leur qualité. La présence de fac-similés, d'échantillons ou de protocoles explicatifs est un attendu indispensable pour évaluer la qualité de l'océrisation et de la transcription de données. Dans la gestion de projets numériques, il faut incorporer cette nécessité de donner les moyens d'évaluer sa propre production.

Même lors de l'utilisation de mêmes vocabulaires contrôlés et de technologies standardisées, les pratiques peuvent différer d'un projet à l'autre. Un travail d'uniformisation sur les fichiers récupérés est alors toujours nécessaire avant leur réutilisation. Autrement dit, l'utilisation de normes et de standards ne suffit pas à supprimer toute ambiguïté ; chaque projet a ses objectifs scientifiques propres et aussi sa propre interprétation des champs et des concepts mis en œuvre par les standards de métadonnées. La façon de remplir des métadonnées aussi génériques que le Dublin Core est laissée à la liberté des éditeurs, d'où des variations énormes dans la teneur de ces métadonnées.

La conclusion s'impose d'elle-même mais est des plus chronophages : il y a nécessité de tout expliciter.

2.5 Bilan

Cet état des lieux nous invite à dresser un premier bilan avec quatre constats.

Du point de vue des données et des métadonnées, il faut pouvoir garantir la présence de métadonnées de base et d'une version brute au format TXT en plus de la version enrichie et des métadonnées plus spécifiques. Les données brutes sont en effet plus facilement réutilisables, mais, comme les données enrichies, elles doivent être accompagnées d'une documentation sur le projet et ses choix éditoriaux.

Il convient de s'assurer que les données produites soient visibles, trouvables et référencées, ainsi que l'ensemble de la documentation les accompagnant.

Il faut pouvoir concevoir, dès la phase initiale du projet, une charte anticipant et précisant toutes les conditions de réutilisation possibles en concertation avec l'ensemble des acteurs concernés (institutions, ayants droit, etc.). Quand un projet implique plusieurs partenaires, ou des partenaires internationaux, qui peuvent être soumis à des cadres légaux différents, il apparaît nécessaire de prévoir dans un document-cadre les difficultés potentielles à cette gestion des données, idéalement dès le tout début du projet.

Concernant la visibilité des données, on pourrait enfin imaginer un lieu ou un outil (vitrine, outil de référencement unique, catalogue) qui permet de lister portails et projets diffusant des corpus d'auteur, avec comme indicateurs la réutilisabilité et la réutilisation effective des données. Il existe le site d'EADH[18] qui propose une liste de projets mais il y manque la possibilité de faire des recherches par type de corpus, par langue, etc., ce qui limite les possibilités de réutilisation.

18 Voir https://eadh.org/projects

3 Penser la réutilisation au début du projet

3.1 Préparer le terrain

Dans la phase d'élaboration du projet, la démarche la plus pertinente est de dresser un état de l'art et de voir s'il existe des données pouvant être réutilisées pour son projet. En parallèle, il est nécessaire de penser à la question de la réutilisabilité des données produites dès le début du projet – et non à la fin, comme c'est souvent le cas. Il faut penser à intégrer cette dimension dans toute réponse à appel à financement pour être certain d'un budget spécifiquement dédié à la réutilisabilité des données. Il est aussi conseillé aux porteurs du projet de se renseigner sur tous les aspects d'ouverture des données liées à la science ouverte[19]. La question de la mémoire du projet et de la sauvegarde des données doit être envisagée dès les prémices du projet.

3.2 Méthodologie et organisation

Trop souvent, les responsables de projet ne sont pas encore forcément sensibilisées à la question du traitement des données. Nous préconisons donc d'avoir, aux côtés du ou de la responsable scientifique, une personne référente « traitement des données » (ingénieur.e, post-doctorant.e, archiviste, doctorant.e, bibliothécaire…) qui peut expliquer les avantages et les limites des différents outils et méthodologies envisagés et qui constitue une véritable force de décision et de proposition au sein de l'équipe.

Il faut également veiller à ce que plusieurs personnes aient accès aux données dans leur intégralité et aux comptes liés aux différents outils, afin d'éviter que tous les accès soient concentrés sur une seule personne. On peut prévoir une ou plusieurs réunions de préparation réunissant l'équipe et permettant de discuter des outils les mieux adaptés aux données et aux compétences des responsables de projet. S'il est recommandé d'utiliser des logiciels libres (et de former les volontaires), il peut néanmoins être contre-productif de l'imposer.

Il apparaît essentiel de s'entendre dès le début avec l'ensemble des membres de l'équipe (en particulier quand celle-ci mobilise différentes disciplines et différentes professions) sur un vocabulaire commun pour éviter les incompréhensions futures : un glossaire est un outil adapté et une des premières tâches à réaliser en début de projet. Il convient également d'adopter un plan de nommage des fichiers. Ces différentes actions permettront de documenter le projet et de s'y repérer facilement à tout moment. Nous préconisons aussi de garder accessibles et d'archiver les documents préparatoires du projet.

19 Voir https://www.ouvrirlascience.fr

3.3 L'organisation interne

Dès le début du projet, il est important :

- d'avoir un lieu de dépôt unique des documents concernant le projet ; en revanche, il faut penser à faire des sauvegardes sur différents supports et à différents endroits.
- de définir les rôles, les droits et les accès de chaque membre du projet.
- de définir une arborescence des fichiers et un plan de nommage partagés par tous les membres de l'équipe (conseil : le nom des fichiers doit contenir la date et la version clairement apparentes). Cette arborescence et ce nommage concernent aussi bien les données que la documentation.

En début de projet, il est conseillé de choisir un outil pour stocker la mémoire du projet et adapté à celui-ci et à ses participants. La mémoire du projet concerne non les données mais la méthodologie, les protocoles, les échanges, etc. De nombreux outils peuvent répondre à ce besoin, tant des logiciels de gestion en gestion électronique de la documentation (GED) que par des solutions en ligne (cloud, Wiki, Agora, Kanboard, Git, GitLab, etc.)

Cette stratégie permet à chaque membre de l'équipe d'être partie prenante de l'ensemble du projet et de son organisation, d'avoir accès à l'ensemble des documents qui le concerne.

3.4 Volet formation

Il est recommandé de proposer une formation qui sensibilise l'ensemble des membres de l'équipe aux enjeux de la réutilisation des données. Elle peut être très courte pour ne pas décourager les participants au projet, auxquels il faut surtout montrer les bénéfices que leur travail pourrait tirer de la réutilisabilité, aussi bien en réutilisant des données qu'en permettant de réutiliser ses propres données. Un bon moyen peut être de s'appuyer sur un cas pratique en montrant aux collègues qu'eux-mêmes s'appuient nécessairement sur des jeux de données préexistants et en les invitant à réfléchir à la manière dont ils pourraient se les réapproprier plus facilement : correction d'un OCR, transcription d'un fac-similé, réutilisation d'autres schémas d'encodage, etc.

3.5 Faire un état des lieux juridique

Un tel état des lieux doit concerner autant les données réutilisées que les données produites et comprendre au moins les aspects suivants :

- identification du statut juridique des données réutilisées ;
- prise en compte des différents types de contenus produits par le projet : images et transcriptions, images sans transcriptions, transcriptions sans images, etc. ;

- intégration de la question de la réutilisabilité des données à la convention entre les différents partenaires, qu'il s'agisse d'institutions, d'individus ou d'ayants droit ;
- résolution des problématiques induites par le RGPD (Règlement général sur la protection des données). En cas de doute, il faut contacter le DPO (Data Protection Officer), aussi appelé DPD (Délégué à la Protection des Données) de votre établissement, qui vous conseillera sur les démarches éventuelles à entreprendre auprès de la CNIL (Commission Informatique et Liberté) et sur les précautions à prendre avant de traiter et de publier vos données ;
- résolution des contraintes liées au respect des droits d'auteurs ;
- état des lieux des possibles embargos pour ne pas se retrouver bloqué au moment de la mise en ligne des données.

3.6 Plan de gestion de données et réutilisabilité

Un plan de gestion des données (PGD) est un document qui envisage la gestion des données dans leur ensemble. Il est de plus en plus demandé d'en rédiger un pour certains appels à projets comme l'ANR ou les appels Collex-Persée. Même sans postuler à un appel à financement, il est pertinent, voire nécessaire d'en élaborer un pour s'assurer de la viabilité du projet. Des outils comme DMP OPIDOR[20] permettent d'en réaliser facilement, de les réactualiser et de les maintenir. Un « bon » PGD doit intégrer et mettre en avant la totalité des thématiques FAIR ; c'est donc un outil à utiliser pour prendre en compte et valoriser les traitements effectués afin de favoriser la réutilisation de données.

Nous proposons que le PGD, qui est un document évolutif s'adaptant aux avancées dans le traitement des données, soit rendu accessible en tant que document administratif mais surtout de document de travail présentant les méthodologies employées. Il permettra à la fois de crédibiliser une démarche de réutilisation de données existantes et de faciliter la réutilisation ultérieure des données produites.

3.7 Deux états des données

Nous recommandons de proposer deux états des données : une version brute (format TXT ou avec un encodage très léger) et une version éditorialisée. Cela facilite grandement la prise en main des données par d'autres. Plusieurs possibilités existent : mettre à disposition les versions non éditorialisées, indiquer une méthode (outils, scripts…) pour obtenir une version brute à partir des données complètes. Il est important de préciser à chaque fois le contexte de production de ces données.

20 Voir https://opidor.fr/planifier

3.8 Choisir ses outils

Il est souvent difficile de déterminer quels outils seront vraiment appropriés au projet, tout en favorisant la réutilisabilité des données par soi ou par d'autres : ils doivent être adaptés aux questions de recherche et aux modes d'organisation de l'équipe, tout en garantissant aux données produites une certaine longévité.

Il est donc conseillé de prendre le temps de réaliser un état de l'art et de s'appuyer sur les retours d'expérience et l'expertise des réseaux que l'on a autour de soi, avec les points d'attention suivants :

- **la licence et les conditions d'utilisation** : il est de loin préférable que l'outil soit libre et ouvert.
- **l'existence d'une communauté** d'utilisateurs et de développeurs autour de l'outil : plus cette communauté est importante et structurée (avec forum, liste de diffusion, documentation et supports...), plus l'outil a des chances d'être durablement maintenu. Il vous sera aussi plus facile de trouver de l'aide en cas de difficulté.

  L'existence d'un réseau d'utilisateurs permet d'évaluer plus facilement, grâce aux retours d'expériences, si l'outil est adapté à votre thématique et à vos données. Il est également essentiel de s'informer sur l'existence de formations, tutoriels ou ateliers plus spécifiques à votre domaine d'expérience, et de la présence de personnes-ressources que vous puissiez contacter.
- **le numéro et la date de version** : un logiciel a un numéro de version qui s'exprime généralement par une suite de nombres séparés par des points (par exemple, Omeka Classic en est à sa version 3.1). Lorsqu'il y a un changement de version, c'est le premier nombre qui change (par exemple, passage de 2.8 à 3.0) ; quand il s'agit d'améliorations ou de débogages d'une version, ce sont les nombres suivants qui évoluent. Quand le premier nombre change très fréquemment, cela peut signifier que l'outil est instable ou qu'il fait l'objet d'une politique commerciale agressive.

  La dernière mise à jour d'un outil ne doit pas être trop ancienne : il doit pouvoir s'adapter aux évolutions des usages et des langages informatiques. Enfin, il faut veiller à la compatibilité des formats entre plusieurs versions d'un même logiciel : il est conseillé de lire avec attention la page des « nouveautés » (appelée fréquemment note de version ou *release*) dans la documentation de la dernière version.
- **les formats** : il faut veiller à ce que l'outil choisi propose, en entrée (import) comme en sortie (export), les principaux formats ouverts (par exemple la suite Libre Office pour les formats bureautiques, un format CSV, XML et/ou JSON pour

les données structurées…) ; pour les données textuelles, on veillera à utiliser l'encodage de caractères UTF-8.

Attention, beaucoup d'outils acceptent de nombreux formats en entrée (import), mais en proposent beaucoup moins en sortie (export) : le principal critère reste que l'export de données depuis l'outil doit pouvoir se faire dans un format ouvert : il faut en effet toujours anticiper et permettre la migration des données vers d'autres outils. Certains outils changent leur liste de formats acceptés lors d'une nouvelle version : c'est pourquoi il est conseillé de faire régulièrement des sauvegardes des données dans un format ouvert.

3.9 Stockage

Pour garantir une meilleure accessibilité des données, il est préférable de ne disposer que d'un seul identifiant pérenne pour chacune d'elles. Il convient donc de stocker ses données dans un endroit unique et sécurisé quitte à les rendre plus visibles en les référençant dans d'autres entrepôts.

Avant de débuter le projet, il est important d'identifier le plus précisément possible le corpus et sa taille afin de mieux orienter les choix de stockage, de sauvegarde et de diffusion. Il est important de souligner qu'un simple stockage sur un disque dur externe ou un espace dans un cloud est une sauvegarde basique mais il faudrait prévoir à intervalle régulière de sauvegarder sur des systèmes plus complets, intégrant indexation et métadonnées, comme les systèmes GED ou les entrepôts de données.

Il est aussi important de bien faire la distinction entre le stockage et la publication des données ; les jeux de données correspondant à ces deux actions sont souvent répartis dans des répertoires et des formats différents. La publication peut notamment être accompagnée d'un *data paper*.

Enfin, il faut œuvrer à la mise à disposition du code ayant permis la réalisation, la diffusion ou l'exploitation de ces données. Plutôt que de mettre à disposition ce code sur le site vitrine du projet, celui-ci peut être stocké dans une forge logicielle, plateforme de diffusion de code source (telle que GitHub ou GitLab). C'est un gage de réutilisation (et d'amélioration) des données, mais aussi un facteur de crédibilisation du travail : on donne le matériau et les moyens de le traiter, ce qui permet la réplicabilité des opérations ainsi que le partage et l'évolution des outils.

3.10 Choix de communication externe

Pour permettre effectivement la réutilisation des données d'un projet, il est important de guider les potentiels futurs utilisateurs au sein de celles-ci. Dans cette optique, expliquer l'organisation des données semble un premier pas nécessaire.

Un *data paper* (ou article de données) peut aussi être un point d'entrée et une source de visibilité très efficace dès la phase initiale du projet puisqu'il constitue une publication scientifique. Déclarer son projet auprès de référentiels disciplinaires (EADH[21]) ou institutionnels est aussi une bonne option.

Un logo "Réutilisez-moi" ou un encadré incitant à la réutilisabilité seraient des initiatives pertinentes en ce sens. Cette signalétique devra mener à la page du site qui détaille les conditions de réutilisabilité des données. Il est important d'informer et de montrer la volonté qu'ont les membres du projet de voir leurs données réutilisées.

21 Voir https://eadh.org/projects

4 Penser la réutilisabilité tout au long du projet

La réutilisabilité doit être réfléchie et préparée à chaque étape du projet, dans tous ses aspects : méthodologique, technique, documentaire, humain. Elle est certes un livrable mais pas seulement : elle est la garantie que les bonnes pratiques liées aux principes FAIR seront respectées et que les compétences acquises, les outils développés et les données produites dans le cadre du projet seront capitalisés.

Cette section du *vademecum* mettra l'accent sur la documentation des données issues du projet : il est en effet essentiel de veiller à conserver la mémoire de chacune de ses phases, y compris les impasses et les revirements. Les multiples données issues du projet peuvent alors être considérées comme un patrimoine, qui doit être géré avec soin : le temps passé à documenter et à archiver chaque produit de la recherche sera largement compensé par les apports méthodologiques et scientifiques de cette pratique.

4.1 Gestion de projet : grands principes

Garantir la réutilisabilité des données tout au long du projet de recherche doit être guidé par quelques grands principes de la gestion de projets, parmi lesquels :

- privilégier la simplicité et la régularité : organiser des réunions régulières et courtes ; se fixer des échéances réalistes et s'y tenir ;
- adapter ses objectifs aux moyens humains, financiers et temporels à disposition ;
- penser à utiliser des outils génériques. Un développement *ad hoc* peut être séduisant au départ car plus facilement adaptable aux exigences scientifiques du moment qu'un outil générique demandant une adaptation des données. Mais, à moyen terme, cela devient contraignant face aux évolutions normales d'un projet de recherche et aux évolutions technologiques. À long terme, cela peut demander une lourde maintenance informatique ;
- assurer la pérennité du suivi du projet : veiller à la transmission et à la formation ; mettre en place un suivi de projet pour faire face aux obstacles identifiés ;
- garder une trace de toutes les étapes du projet et des documents produits, du processus de constitution du corpus à la mise en œuvre du protocole éditorial, puis mettre à disposition ces documents essentiels à l'intelligibilité des données ;
- penser à l'après projet : envisager les différents futurs du projet (l'arrêt, la prolongation, la transformation, l'association à d'autres projets…).

4.2 En pratique

Permettre la réutilisation de données produites au cours du projet demande une organisation interne rigoureuse, mais aussi la mise en place d'une communication externe régulière pour informer la communauté de la disponibilité des jeux de données.

4.2.1 L'organisation interne

Fluidifier le partage d'informations entre membres de l'équipe est un des points qui mérite le plus d'attention.

Tout au long du projet, il est important :

- de faire régulièrement des points « réutilisabilité » ;
- de garder une trace de tous les échanges et de déposer les comptes-rendus dans un espace dédié ; ceux-ci doivent être simples, brefs et synthétiques avec une liste des personnes présentes, la date, le relevé de décisions, si possible au format TXT. Pourquoi le format TXT ? Parce que c'est un format brut, indépendant de tout environnement logiciel, dont la stabilité, l'accessibilité et la pérennité sont donc garanties !
- de faire du suivi de versions lorsque les documents ont de multiples évolutions, mais aussi de faire le tri et de supprimer les versions peu significatives : ne pas garder l'inutile, comme, par exemple, les versions qui ne reflètent que des changements de mise en forme ;
- de décider d'une politique de dépôt des données du projet : qui dépose ? à quelle fréquence ? à quel moment les données en cours de constitution passent-elle du GED où sont stockées les versions de travail, à l'entrepôt où elles seront exposées ? À quel moment sont-elles rendues publiques ?

4.2.2 Quels outils ?

Une veille technologique (distincte de « l'état de l'art » qui est une photographie à l'instant *t* du domaine) doit être conduite et complétée par une veille sur les évolutions technologiques et les formations disponibles, car il est fort probable que les membres de l'équipe soient amenés à faire évoluer leurs compétences.

Le choix des outils dépend des caractéristiques du projet, de sa temporalité, de l'équipe, mais aussi des solutions mises à disposition par les institutions.

Parmi les entrepôts et gestionnaires de fichiers existant, on peut citer :

- des entrepôts (Didómena[22], Nakala[23], Dataverse[24], Zenodo[25]…) permettant l'exposition de données, souvent « froides » ou qui ne font pas l'objet de modifications régulières ;
- des gestionnaire de fichiers (Sharedocs[26] mis en œuvre par Huma-Num, My Core[27] au CNRS ou encore d'autre cloud institutionnels) qui permettent le stockage et le partage de fichiers au sein de groupes de travail. Certains offrent aussi des fonctionnalités d'écriture collaborative.
- des outils de versionnage (GitLab[28]…), plateformes comprenant aussi des fonctionnalités de suivi des bugs, des espaces de discussions et de décisions, des wikis, etc.

4.2.3 La communication externe

La communication externe permet de rendre compte de l'évolution du projet, voire de formaliser certains choix méthodologiques. Il est donc intéressant de se doter assez tôt d'outils de communication externe permettant de tenir la communauté informée de l'existence du projet, de ses développements et des données disponibles.

La communication externe peut prendre la forme de blogs (notamment les carnets Hypotheses.org[29]), de lettres hebdomadaires, d'actualités sur un site, etc. Il ne faut pas hésiter à exposer les problèmes théoriques et techniques qui se posent.

La publication des données sur un site internet au fur et à mesure et de manière régulière permet de diffuser des données en cours de constitution, surtout dans le cadre de projets où elles évoluent constamment et où le dépôt finalisé dans un entrepôt apparaît moins pertinent.

Sur la page vitrine de votre site, indiquez clairement les lieux où trouver les différents types de données : fichiers XML, fichiers images, documentation, code source…

4.3 Méthodologie

Produire une documentation de ses données et de ses méthodes, est essentiel pour permettre la réutilisation des données. Cette documentation doit être enrichie tout au long du projet.

22 Voir https://didomena.ehess.fr
23 Voir https://www.nakala.fr
24 Voir https://dataverse.org Il est annoncé une instance nationale pour début 2022.
25 Voir https://zenodo.org
26 Voir https://sharedocs.huma-num.fr
27 Voir https://mycore.core-cloud.net
28 Voir https://about.gitlab.com
29 Voir https://hypotheses.org

4.3.1 Documentation des données

Pour que les données soient réutilisables, la documentation doit être de qualité et concerner absolument tous les aspects et contenus du projet : corpus, codes, scripts, protocoles... Tous les traitements effectués sur les données doivent être documentés (conversions, OCR, enrichissements, etc.). L'organisation et la structuration des données doivent aussi être expliquées et décrites. Lors de la publication des données, on pourra également joindre tous les documents nécessaires à leur intelligibilité (modèle de données, schéma d'éditorialisation, manuel d'encodage, etc.).

A/ La documentation des données doit se faire :

- à travers des normes et standards de métadonnées (Dublin Core, Data Documentation Initiative, EAD, Mets, MODS, Exif, IPTC, etc.). Les choix de métadonnées peuvent notamment apparaître dans une section dédiée du PGD ;
- à travers des documents rédigés qui expliquent dans un langage naturel l'historique du projet, l'élaboration du jeu de données et les choix éditoriaux. Il peut exister des formats spécifiques pour produire cette documentation : il est utile de se renseigner et/ou de se former pour les utiliser, mais, à défaut, un simple document texte (TXT, voir ci-dessus !) est aussi un bon choix.

Ces deux options ne s'excluent pas l'une l'autre, mais sont complémentaires.

Dans l'idéal, il faudrait aussi générer la documentation en anglais, notamment dans le cadre de projets regroupant des partenaires internationaux.

La documentation des données peut aussi passer par la rédaction d'un *data paper*, qui est une excellente façon de rendre compte de la démarche scientifique ayant présidé à la constitution des données tout en assurant leur visibilité.

B/ La documentation du traitement des données peut prendre des formes variées : schéma au format RNG pour des données encodées en XML ; fichiers modèles ; fichiers de recollements, tableaux de correspondances pour les numérisations ; modèles de données pour les bases de données relationnelles[30], etc.

Dans le cadre d'éditions encodées en XML-TEI, un document ODD constitue un outil de référence pour les membres du projet, tout en fournissant la documentation la plus complète possible pour des personnes extérieures. Le principal intérêt du format ODD est en effet de réunir dans un même document le schéma d'encodage, avec les différentes balises

30 Etalab, département de la direction interministérielle du numérique (DINUM) qui a pour mission de coordonner la conception et la mise en œuvre de la stratégie de l'État dans le domaine de la donnée, propose ainsi un guide pour décrire son modèle de données : https://guides.etalab.gouv.fr/qualite/documenter-les-donnees/#description-du-modele-de-donnees

utilisées et les règles propres à leur utilisation, et une documentation structurée et rédigée concernant la constitution du corpus, les enjeux du projet et les exploitations possibles, le protocole éditorial adopté, les choix d'encodage, en insistant sur leurs particularités[31].

La documentation du code informatique est également essentielle, à plus forte raison quand plusieurs développeurs interviennent. Pour cela, l'utilisation d'un fichier Readme est nécessaire. C'est un fichier donnant des informations sur les autres fichiers contenus dans le même répertoire. Généralement au format texte ou markdown, on y retrouve en principe les instructions d'installation et d'exploitation, l'explication de l'arborescence du répertoire, la liste des autres fichiers avec un descriptif de leur contenu, des liens vers les auteurs du projet et la licence applicable.

4.3.2 Contrôler et évaluer la qualité des données

Pour que la réutilisabilité des données soit possible, il faut se donner les moyens de garantir leur qualité. Les réutilisateurs potentiels doivent pouvoir :

- vérifier les transcriptions : certains projets proposent par exemple systématiquement le fac-similé accessible à côté de la transcription ;
- comprendre comment les données textuelles ont été produites : quels choix d'encodage et de structuration ont été faits, pour quelles raisons ? Documenter les options retenues mais aussi celles qui ont été écartées est donc extrêmement utile pour la gestion du projet lui-même, comme pour sa réutilisation ;
- pouvoir lire, vérifier et modifier le code informatique : il faut avoir accès à une documentation systématique et rédigée de ce dernier. Pour la réutilisation du code, une documentation interne, directement dans le code et pas à pas, est indispensable.

L'écriture exécutable est une méthode adéquate pour rendre son code réutilisable. JupyterLab[32] fournit ainsi un environnement dans lequel des carnets Jupyter notebook peuvent s'exécuter. Ce type d'écriture a pour intérêt de réunir documentation, codes, scripts, résultats et visualisations dans un même document, tout en fournissant un environnement intégrant des langages de programmation et des bibliothèques associées.

31 Pour en savoir plus, cf. présentation de Lou Burnard, « Comment maîtriser le tigre TEI », 2018 (https://cahier.hypotheses.org/files/2018/08/ODD-diapos.pdf) et Jean-Baptiste Camps, « Structuration des données et des documents : balisage XML », 2018, (https://halshs.archives-ouvertes.fr/cel-01706530/file/00_Syllabus_20151020.pdf).
32 Voir https://jupyter.org

5 Bilan

Ce travail collectif nous a permis de faire les deux constats suivants :

- les données réutilisables restent peu nombreuses et sont souvent difficiles à trouver : les moyens et les outils permettant de les retrouver restent à perfectionner ;
- la pérennité de l'accès aux données reste un problème fréquent, surtout quand elles n'ont été diffusées que par l'intermédiaire d'un site internet. Après la fin du projet, les sites n'étant plus régulièrement (ou plus du tout) maintenus, les données deviennent inaccessibles. Il faut alors prévoir des solutions pour que le site soit gelé avec uniquement des mises à jour de sécurité, permettant de maintenir une accessibilité à des données datées. Le dépôt dans un entrepôt s'avère, à terme, une des solutions les plus satisfaisantes pour ouvrir les possibilités de réutilisation et garantir une préservation à moyen terme.

Nous espérons que la lecture de ce *vademecum*, vous permettra de travailler, pendant toute la conduite de nos projets, à la réutilisabilité de vos données : c'est une garantie de qualité et de visibilité.

Et si nous réutilisons nous-même des données :

- pensons à prévenir l'équipe productrice, qui sera sans doute ravie de savoir son travail prolongé et enrichi par d'autres ;
- reconnaissons toujours nos dettes : citons systématiquement les producteurs et... pensons à les remercier !

Annexes

A Check-List « réutilisabilité »

Cette check-list est inspirée de certains points que l'on doit retrouver dans un plan de gestion de données (PGD).

- Existence de données déjà existantes que le projet pourrait réutiliser ?
- Ces données sont-elles accessibles ? Juridiquement ? Techniquement ? Besoin de conventions ?
- Présence d'un référent « traitement des données » pour le projet ?
- Production d'une liste des données qui seront produites ? Données des sources de la recherche, données textuelles (ou autres), facsimilés et transcriptions ? Données premières, secondaires, tertiaires ? Schéma de métadonnées, schéma d'encodage ; code source, script associé ?
- Dans cette liste, les données à rendre réutilisables sont-elles identifiées ? Types, niveaux, versions ?
- Le statut juridique des données réutilisables est-il prévu (choix de la licence) ? Licences (prise en compte des différents cas de figure - institution, ayants droits, etc.), délais, embargos ?
- Publication d'indicateurs de qualité des données ? Fac-similés pour transcriptions, taux d'erreur, relecture, protocole de validation ?
- Création d'un document de suivi de projet récapitulant les décisions prises au cours de celui-ci ? Choix éditoriaux, choix méthodologiques, réajustements, revirements ?
- Documentation (scientifique, méthodologique, technique) prévue ? Où sera-t-elle stockée ? Par qui ? Qui y a aura accès ?
- Des réutilisations ont-elles été envisagées ?
- Les moyens de rendre visible ces données ont-t-ils été anticipés ? Référencement, « trouvabilité », « citabilité » : entrepôt, site web, git, etc.
- Est-il prévu d'indiquer les différents lieux où sont déposées les données sur la page d'accueil du site ?
- Où les données et les métadonnées seront-elles stockées tout au long du projet ? Après le projet ?
- L'archivage des données est-il prévu ?

B Trucs & astuces

Permettre de retourner à la version brute

Les textes bruts sont souvent ceux qui sont le plus aisément réutilisables ; or l'encodage a souvent lieu en même temps que les transcriptions (l'expérience montre d'ailleurs qu'il vaut mieux procéder ainsi plutôt que de transcrire « au kilomètre » pour revenir ensuite sur les transcriptions). Il est donc utile, y compris dans le cadre du projet, de prévoir un script qui enlève toutes les balises (TEI ou autres).

Un script (Python, XSLT ou autre) contenant la regex <[^<]*> suffit pour enlever les balises *a posteriori*.

Il est important de conserver cette « version brute » pour toutes les données qui ont subi des traitements. Par exemple, pour les images ou les documents multimédias, on peut mettre en ligne des formats compressés adaptés au support de diffusion, mais il faudra conserver les formats en haute ou très haute définition et les déposer sur un entrepôt.

Déposer son corpus de textes TEI dans un entrepôt

Lorsque l'on dépose des données dans un entrepôt, les données issues du corpus peuvent se retrouver isolées les unes des autres. C'est typiquement le cas si on utilise un fichier TEI avec un <teiCorpus>[33], qui inclut d'autres fichiers XML (avec une balise TEI appelant chaque fichier). Le <teiCorpus> contiendra bien un lien vers les autres fichiers mais ceux-ci se retrouveront « orphelins ». Pour pallier ce problème, différentes solutions sont possibles :

- faire un seul dépôt avec l'ensemble des textes regroupés au sein d'un même fichier[34] ;
- faire le lien entre le fichier contenant le <teiCorpus> et le fichier TEI dans les métadonnées (champ « relation » du Dublin Core par exemple) ;
- dans un entrepôt comme Nakala, où plusieurs fichiers peuvent être associés à une donnée, déposer avec chaque fichier TEI le fichier contenant le <teiCorpus>.

Déposer son corpus de fichiers d'un Omeka dans un entrepôt de données

Un outil comme Omeka (Classic ou S) permet l'export des métadonnées tout au long du projet, soit de manière manuelle (en CSV par exemple), soit *via* un moissonnage OAI-

33 Un « TEI Corpus » est un fichier qui contient la totalité d'un corpus encodé selon la TEI, comprenant un seul en-tête de corpus et un ou plusieurs éléments TEI dont chacun contient un seul en-tête textuel et un texte. Source : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-teiCorpus.html

34 Les fichiers TEI ne sont plus appelés par un <xi :include> mais les textes sont directement encodés dans un <teiCorpus> global.

PMH. Concernant les métadonnées, un export manuel peut servir de base à un dépôt dans un entrepôt. Il faut toutefois veiller à reproduire, lors du dépôt, la correspondance entre les métadonnées et les données (fichiers sous Omeka). Concernant les données, il faudra récupérer l'ensemble des fichiers. Malheureusement, dans Omeka Classic notamment, ils sont stockés dans plusieurs formats (original, réduits, vignette) et renommés automatiquement : une requête en base de données pour obtenir le nom original du fichier est indispensable.

Assurer le relais entre les acteurs : le « kit de départ »

Pour garantir le bon suivi du projet malgré les changements de personnel, il faudrait instaurer la conception d'un « kit de départ ». Toute personne en CDD, stage ou prestation de service doit laisser à son successeur de quoi prendre le relais sans perte de temps et d'informations. Il s'agit de communiquer clairement les informations de suivi du projet, les accès aux différentes ressources, l'arborescence des organigrammes, etc.

Un bon kit de départ est celui que l'on aurait souhaité trouver à son arrivée.

Prise de notes et comptes-rendus

Lors de l'établissement des documents et des conventions de travail, pour la fluidité dans la transmission d'informations, chaque membre de l'équipe prend des notes et les place dans un dossier dédié partagé. C'est particulièrement utile pour les comptes-rendus de réunion. Mais il ne faut pas oublier de mandater quelqu'un pour faire un compte-rendu synthétique avec, surtout, une liste claire des décisions (la rubrique « relevé de décisions » est alors obligatoire).

Publier

Publier sur la façon de produire ses données (sous la forme d'un *data paper* dans une revue dédiée ou d'un article plus traditionnel) permet de documenter l'intégralité des processus du projet de façon rigoureuse, aussi bien en ce qui concerne les choix scientifiques que les choix techniques, en mettant en évidence le lien indissoluble de ces choix. Pour les chercheurs, cette pratique permet en outre de rendre visible cette part essentielle, mais souvent peu valorisée, de leur travail.

Choisir un format

Privilégier les formats textes (TXT, XML, CSV, etc.) qui permettent des conversions lors du changement de logiciels ou de plates-formes. Il faut aussi oser changer de format quand

cela s'avère nécessaire : d'un format propriétaire à un format ouvert, d'un XML « maison » à l'XML-TEI par exemple.

Il ne faut pas oublier le format d'encodage des caractères. Unicode (UT8) est indispensable.

Versionner !

Versionner rigoureusement les différentes étapes du projet (élaboration des données, rédaction des différents documents) permet de revenir à l'étape précédente et peut même s'avérer crucial en cas d'erreur ou de revirement. Pour cela, il faut conserver les différentes versions d'un code source, les étapes dans le traitement des données, les différents états de la documentation, etc.

Concernant les développements informatiques, des outils sont déjà largement utilisés (Git). En revanche, pour d'autres types de données, il n'y a pas toujours de solutions évidentes. Mais pourtant elles existent : HAL permet de conserver différentes versions d'une même publication, des entrepôts tels que Nakala et Dataverse aussi, tout comme certaines plateformes de rédaction collaborative (Overleaf pour la rédaction de documents en LaTeX, Etherpad pour les fameux Framapad ou encore Stylo, l'éditeur de texte sémantique proposé comme service par Huma-Num).

Un dernier conseil est celui de repérer et de nommer certaines versions « clefs » qui marquent une étape spécifique dans le déroulement du projet.

C Glossaire des notions et des outils

Ce glossaire n'a pas de prétention exhaustive. Il explicite uniquement les termes utilisés dans le *vade-mecum*. Les termes soulignés renvoient à une entrée du glossaire. Pour avoir un glossaire plus complet sur les humanités numériques, vous pouvez consulter utilement celui édité par l'initiative Digit Hum et qui nous a servi de base pour la constitution de celui-ci[35].

- **Accaparement** : Voir *Scraping*.

- **API (*Application Programming Interface*)** : Une API (ou interface de programmation applicative) permet à deux logiciels de communiquer entre eux. Un programme va solliciter des services auprès d'un autre programme avec des appels de fonction : des codes rassemblés dans une API. La syntaxe requise par l'API est décrite dans la documentation de l'application fournisseur, qui va spécifier comment des programmes consommateurs peuvent se servir de ses fonctionnalités. L'API permet donc de réutiliser des briques de fonctionnalités fournies par d'autres logiciels[36].

- **CMS (*Content System Management*) :** Le CMS (ou système de gestion de contenu) est un système d'interfaces qui permet de gérer la conception et la gestion d'un site sans avoir beaucoup de connaissances en informatique. Il ne peut toutefois jouer le rôle d'une base de données ou d'un éditeur numérique pour générer des éditions scientifiques enrichies. Parmi les CSM open source les plus connus, on peut citer Wordpress, Drupal, Omeka.

- **CSV (*Comma Separated Values*)** : Le CSV est un format informatique stockant des données sous forme de valeurs séparées par un code ; le plus souvent le séparateur est la virgule ou le point-virgule. La représentation usuelle de ces données mémorisées en CSV est le tableau. Ce format ne permet pas d'enrichissement typographique (gras, italique, etc.), il conserve du texte brut. Attention, les contenus doivent être codés en Unicode[37].

- **DDI (*Data Documentation Initiative*)** : C'est un standard de documentation technique pour décrire et conserver les informations statistiques et plus globalement les données d'enquêtes en sciences humaines et sociales.

- ***Data paper*** : Le *data paper* (*data article*, *data descriptor*) est une publication qui décrit un

35 Voir https://digithum.huma-num.fr/ressources/glossaire
36 Source : « Glossaire des termes techniques EMAN », Site *EMAN (Édition de Manuscrits et d'Archives Numériques)*, https://eman-archives.org/EMAN/glossaire-eman. Consulté le 20/01/2022 sur la plateforme EMAN.
37 Source : *Idem*.

jeu de données scientifiques, notamment à l'aide d'informations structurées appelées métadonnées. Contrairement aux articles de recherches classiques, les *data papers* fournissent une voie formalisée au partage des données plutôt que tester des hypothèses ou présenter de nouvelles analyses. On trouve dans son contenu une introduction sur l'intérêt du jeu de données, la description des données en elles-mêmes, la méthodologie employée pour l'obtention de données, des remerciements et les références nécessaires[38].

- **Dataverse** : Dataverse est un logiciel open source de création et de gestion d'entrepôts de données. Un identifiant numérique DOI est attribué à tous jeux de données. L'accès à ces données peut être ouvert à tous ou restreint selon des conditions d'utilisations mentionnées. Dataverse s'appuie sur des normes permettant l'échange de métadonnées, leur indexation par les moteurs de recherche et facilite la mise en réseau de portails scientifiques pluridisciplinaires. La plateforme accueille tous formats de fichiers, de préférence ouverts et standards pour en faciliter le partage et la réutilisation. S'inscrivant dans le courant de la Science Ouverte, Dataverse fournit une réponse adaptée pour gérer les données produites par la recherche académique.

- **DOI (*Digital Object Identifier system*) :** Le DOI est un mécanisme d'identification de ressources possiblement numériques mais pas seulement. (ex : films, articles scientifiques, personnes ou autres). Il facilite la gestion, car chaque DOI est unique pour chaque ressource, tout en associant des métadonnées à l'identifiant. Il constitue une alternative à l'instabilité des URL.

- **Dublin Core** : Le Dublin Core est un standard de métadonnées permettant de décrire tout type de ressources numériques. Il existe 15 champs généralistes, répétables et facultatifs. Le site de la BnF récapitule ces différents champs ainsi que les différentes possibilités d'utilisation ou de raffinements[39]**.**

- **EAD (*Encoded Archival Description)*** : La description archivistique encodée est un standard d'encodage international des documents archivistiques basé sur le langage XML, maintenu par la Société des archivistes américains en partenariat avec la Bibliothèque du Congrès sur le site de cette dernière[40]**.**

- **Écritures exécutables** : Voir Jupyter notebook.

- **Entrepôt de données :** Un entrepôt de données de recherche (*Research Data Repository* ou *Data Repository*) est une banque de données destinée à accueillir, conserver, rendre visibles et accessibles des données. Son rôle est de permettre le dépôt ou la collecte de données, leur description, leur accès, et leur partage en vue de leur réutilisation. Chaque

38 Source : *Data papers et data journals : FICHE SYNTHÉTIQUE***,** Publié le 10/02/2017 | Mis à jour le 09/07/2018 | DOI : 10.13143/2wcb-fw52.
39 Voir https://www.bnf.fr/fr/dublin-core
40 Voir https://www.loc.gov/ead

entrepôt dispose généralement d'une politique de dépôt, de description et de diffusion des données.[41]**.**

- **Format libre :** Le format libre est un format qui n'est la propriété de personne et qui peut donc être exploitable par tous (ex. PNG, JPEG)[42].

- **Format ouvert :** Un format ouvert évoque tout protocole de communication ou d'échange ou tout format de données interopérable dont les spécifications techniques sont publiques et sans restriction d'accès ni de mise en œuvre[43].

- **GED (Gestion Électronique des Documents)** : La GED est un procédé faisant intervenir des moyens électroniques – typiquement des logiciels et/ou des matériels – pour prendre en charge la gestion des documents, à savoir les opérations et actions destinées à traiter et à exploiter les documents, par exemple la capture, l'acquisition, la numérisation, la validation, la diffusion, le classement, l'indexation, l'archivage, etc. Sharedocs, outil proposé par Huma-Num, ou MyCore, utilisé au CNRS, sont tous deux des outils de GED.[44]

- **Git :** Logiciel de gestion de versions de code source. Projet *open source*, c'est un des plus utilisés au monde.

- **Gitlab :** Plateforme de développement collaborative *open source*. Se basant sur les fonctionnalités du logiciel Git, elle permet de piloter des dépôts de code source et de gérer leurs différentes versions. Son usage est particulièrement indiqué pour les développeurs qui souhaitent disposer d'un outil réactif et accessible.

- ***Handle System* :** Mécanisme d'attribution d'identifiants pérennes pour des objets numériques ou d'autres ressources internet.

- **HTR (*Handwritten Text Recognition*)** : Cet acronyme désigne les procédés informatiques pour la traduction d'images de textes manuscrits en fichiers de texte.

- **IIIF (*International Image Interoperability Framework*)** : IIIF désigne à la fois une communauté et un ensemble de spécifications techniques dont l'objectif est de créer un cadre technique commun grâce auquel les bibliothèques numériques peuvent diffuser leurs images de manière standardisée sur le Web, afin de les rendre consultables, manipulables et annotables par n'importe quelle application ou logiciel compatible[45].

41 Source : Dedieu, L. ; Barale, M. 2020. *Déposer des données dans un entrepôt, en 6 points*. Montpellier (FRA) : CIRAD, 4 p. https://doi.org/10.18167/coopist/0070
42 Voir https://digithum.huma-num.fr/ressources/glossaire
43 Source : https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000001339538
44 Source : https://www.ged.fr/definition-ged
45 Voir https://iiif.biblissima.fr

- **Jupyter notebook :** Application web open-source permettant de créer et de partager des documents qui contiennent du code exécutable directement depuis l'interface, des équations, des visualisations de données et de la documentation écrite. Ces carnets peuvent être utiliser pour nettoyer et transformer les données, faire des simulations numériques, des modèles statistiques, des visualisations de données ou encore de l'apprentissage automatique, un des champs d'études de l'intelligence artificielle[46].

- **LaTeX :** LaTeX est un logiciel de balisage et de composition de documents spécialement créé pour l'édition de documents scientifiques.

- **Logiciel libre** : Un logiciel libre est un logiciel dont l'utilisation, l'étude, la modification et la duplication par autrui en vue de sa diffusion sont permises, techniquement et juridiquement, ceci afin de garantir certaines libertés induites, dont le contrôle du programme par l'utilisateur et la possibilité de partage entre individus[47].

- **Loi pour une République numérique** : La loi Lemaire pour une République numérique du 7 octobre 2016, publiée au Journal officiel du 8 octobre 2016, vise à favoriser l'ouverture et la circulation des données et du savoir, à garantir un environnement numérique ouvert et respectueux de la vie privée des internautes et à faciliter l'accès des citoyens au numérique[48].

- **Loi Valter** : La loi Valter est la loi du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public[49]. Associée à la loi Lemaire (loi pour une République Numérique), elle a élargi le champ d'application du droit de la réutilisation en y incluant les établissements et services culturels et les établissements d'enseignement et de recherche, qui en étaient auparavant exclus. Les dispositions réglementaires relatives à la réutilisation des données de recherche, considérées au même titre que les informations issues des administrations publiques, sont définies dans le Code des relations entre le public et l'administration, livre III, titre II[50].
- **Métadonnées :** Données structurées décrivant une ressource ou une autre donnée. Les métadonnées servent à référencer, identifier et partager correctement un document. Elles permettent la description et le traitement des ressources numériques. Elles sont généralement standardisées, se plaçant à l'extérieur ou en entête du texte ou du document qu'elles décrivent. On distingue plusieurs types de métadonnées, descriptives (EAD, Dublin Core, MODS), techniques (EXIF, MIX-NISO, etc.), de structure (ALTO, METS, TEI).

46 Voir https://www.arthurperret.fr/du-notebook-au-bloc-code.html et https://hnlab.huma-num.fr/blog/2021/05/26/callisto-un-demonstrateur-jupyter
47 Source : https://fr.wikipedia.org/wiki/Logiciel_libre
48 Voir https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829
49 Voir https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031701525
50 Voir Bruno Ricard, « Le nouveau régime juridique de la réutilisation des informations publiques », carnet *Droit(s) des archives*, 16 mai 2017 : https://siafdroit.hypotheses.org/659

- **Nakala :** Nakala est un entrepôt de données développé par la TGIR Huma-Num.

- **OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*)** : C'est un protocole informatique permettant d'échanger des métadonnées : il permet donc de constituer des entrepôts de données centralisés, regroupant des métadonnées issues de différentes sources, récupérées par "moissonnage" (*harvesting*). Le protocole permet donc un meilleur signalement des données en rendant possible le moissonnage des métadonnées par des services tiers (type Isidore Science par exemple) ; on peut aussi interroger directement un entrepôt de données OAI-PMH en utilisant un langage de requête spécifique.

- **OCR (*Optical Character Recognition*)** : Cela désigne les procédés informatiques pour la traduction d'images de textes imprimés ou dactylographiés en fichiers de texte.

- **ODD (*One Document Does it all*)** : La personnalisation ODD est un format de spécification compatible TEI permettant de personnaliser son usage de la TEI. Son principal intérêt réside dans le fait qu'elle combine dans un même document la documentation rédigée de l'édition et le schéma de structuration des données[51].

- **Omeka** : Logiciel de gestion de bibliothèque numérique mis à disposition sous licence libre. De conception modulaire, l'outil permet à chaque site d'adapter les fonctionnalités proposées à l'aide de plugins et de thèmes. L'outil est développé par le Roy Rosenzweig Center for History and New Media (CHNM) de l'Université George Mason (USA) qui est aussi à l'origine du logiciel de gestion bibliographique Zotero.

- **PGD (Plan de Gestion de Données)** : Le plan de gestion des données, aussi appelé DMP (*Data Management Plan*) se présente sous forme d'un document structuré en rubriques. Il a pour objectif de synthétiser la description et l'évolution des jeux de données de votre projet de recherche. Il prépare le partage, la réutilisation et la pérennisation des données[52].

- **Plateforme** : Une plateforme numérique est un service qui donne accès à des contenus ou services qu'elle n'a pas créés elle-même mais dont elle facilite l'exposition et l'usage. Par exemple, Open Edition Books est une plateforme qui donne accès à des livres publiés par différentes maisons d'édition ; il n'est pas lui-même éditeur des ouvrages mais en permet l'accès en ligne.

- **Portail** : Un site portail est un site offrant un accès unique à plusieurs services différents, voire une porte d'entrée unique vers plusieurs sites différents. Il est parfois difficile de le distinguer d'une plateforme. Par exemple, Open Edition est un *portail* qui donne accès à plusieurs plateformes comme Open Edition Journals, Open Edition Books, Hypotheses et Calenda.

51 Voir http://www.tei-c.org/Guidelines/Customization/odds.xml
52 Voir https://doranum.fr/plan-gestion-donnees-dmp

- **RDF** : voir Triplet RDF.

- **README** : Un fichier *readme* est généralement un fichier texte contenant des informations sur les autres fichiers du même répertoire. Son contenu varie mais inclut d'ordinaire des instructions d'exploitation, une liste des noms et utilités des autres fichiers, des informations sur la personne les ayant créés, voire la licence applicable.

- **Référencement** : Le référencement est, sur le web, l'action de référencer, c'est-à-dire d'indexer toutes les pages web présentes, en faisant un lien d'une page vers une ressource, généralement un moteur de recherche. Aujourd'hui, le référencement consiste surtout à améliorer la place d'un site dans les résultats afin d'être le plus consulté possible.

- **Regex :** Les *regex*, ou expressions régulières, sont utilisées, quand on traite des données textuelles, pour repérer (et éventuellement modifier) des chaînes de caractères. Elles se présentent elles-mêmes sous la forme d'une chaîne de caractères qui, suivant une syntaxe précise, décrit un *pattern* à rechercher dans un document textuel. Par exemple, on peut, à l'aide d'une expression régulière, rechercher dans un texte n'importe quelle séquence de 4 chiffres, n'importe quelle une adresse e-mail, une suite de trois mots dont le second commence par la lettre b, les différentes variantes orthographiques du mot "yaourt" (yoghourt, yogourt, ...), etc. Elles s'utilisent au sein de différents langages informatiques, mais aussi à travers la fonction rechercher/remplacer d'un simple éditeur de texte.

- **RGPD (Règlement général sur la protection des données)** : Émanant du Parlement Européen, il vise à protéger les données personnelles, en donnant aux personnes physiques le contrôle sur les données les concernant (droit d'accès, de rectification, de destruction notamment)[53].

- **Schéma de données** : Un schéma de données est un modèle qui permet de décrire de manière précise et univoque les différents champs et valeurs possibles qui composent un jeu de données[54].

- **Scraping** : Script ou programme qui permet d'extraire du contenu de sites Web pour le transformer ou l'intégrer dans un autre contexte. Le terme a mauvaise presse car souvent l'opération se fait sans tenir compte des droits. L'extraction des métadonnées pour par exemple le référencement dans un site portail est plus souvent dénommée *harvesting* (moissonnage).

- **Sharedocs** : Sharedocs est un système de gestion de documents (GED), ou gestionnaire de fichiers, proposé par la TGIR Huma-Num ; il est conçu pour faciliter l'échange, le

53 On peut le consulter à la page suivante : https://www.cnil.fr/fr/reglement-europeen-protection-donnees
54 Source : Guide Etalab *Préparer les données à l'ouverture et à la circulation*, https://guides.etalab.gouv.fr/qualite/preparer-le-jeu-de-donnees/#la-structure-du-jeu-de-donnees

traitement et l'organisation commune des données d'un projet[55].

- **TEI (*Text Encoding Initiative*)** : Consortium qui développe et maintient collectivement un standard XML pour la représentation des textes numériques[56].

- **Triplets RDF** : Modèle simplifié de description de données dont le principe de base consiste à transformer l'information des ressources afin qu'elles puissent être lisibles par les machines et permettre, par conséquent, la création de liens à partir des valeurs des relations. Sa « grammaire » est constituée de triplets de trois éléments : sujet, prédicat et objet. Les données RDF sont stockées dans un *triple store*.

- **Vocabulaire contrôlé :** Un vocabulaire contrôlé est un ensemble organisé de mots et d'expressions utilisés pour indexer du contenu et/ou le retrouver par navigation ou recherche. Typiquement, il inclut des termes préférentiels et leurs variantes et opère dans un périmètre défini ou décrit un domaine spécifique. L'utilisation d'un vocabulaire contrôlé améliore l'interopérabilité entre différents systèmes comme les moissonneurs, les systèmes d'information de recherche (CRIS), les entrepôts de données et les éditeurs[57].

- **Web sémantique** : Le Web sémantique, appelé aussi Web de données, est le Web permettant d'échanger et d'utiliser des données, de publier et de lier des bases de données sur le Web. Succédant au Web documentaire, il s'appuie sur un standard du Web, l'URI (*Uniform Resource Identifier*), qui identifie une ressource. Le modèle de données RDF, également standard du Web sémantique, permet quant à lui de décrire, représenter et relier des données.

- **Wiki :** Un wiki est une application web qui permet la création, la modification et l'illustration collaboratives de pages à l'intérieur d'un site web. Il utilise un langage de balisage et son contenu est modifiable au moyen d'un navigateur web[58].

- **XML (eXtended Markup Language)** : C'est un langage informatique de balisage qui permet de structurer et d'échanger des données. Il fait partie des recommandations du W3C[59].

55 Pour en savoir plus : https://documentation.huma-num.fr/sharedocs-stockage
56 Site officiel : https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html
57 Voir https://www.coar-repositories.org/files/coar-cv-infog-f_27052042-3.pdf
58 Voir https://fr.wikipedia.org/wiki/Wiki
59 Voir https://www.w3.org/standards/xml/core

D Bibliographie

Les références internet ont été consultées le 18 janvier 2022.

1. Science ouverte

1.1. Généralités

BECARD N. et al. (2017). *Ouverture des données de recherche. Guide d'analyse du cadre juridique en France, v.2*. INRA, MESRI. En ligne : https://www.enssib.fr/bibliotheque-numerique/documents/68091-ouverture-des-donnees-de-recherche-guide-d-analyse-du-cadre-juridique-en-france.pdf

CoopIST (s. d.). « Gérer les données de la recherche. » Site *Coopérer en information scientifique et technique*. En ligne : https://coop-ist.cirad.fr/gerer-des-donnees

InIST-CNRS (2020). *Parcours interactif sur la gestion des données de la recherche*. Site *Doranum*. En ligne : https://doranum.fr/enjeux-benefices/parcours-interactif-sur-la-gestion-des-donnees-de-la-recherche

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (juil. 2021). *Deuxième Plan national pour la science ouverte. Généraliser la science ouverte en France. 2021-2024*. En ligne : https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte

MSH Lorraine (2021). Guide *Ouvrir ses données*. CNRS-Université de Lorraine. En ligne : https://msh-lorraine.fr/nos-services/se-familiariser-avec-la-science-ouverte

1.2. Plans de gestion de données

Doranum (s.d.), *Plan de gestion de données* [liste de ressources]. En ligne : https://doranum.fr/categories/dmp.
Voir en particulier : Inist-CNRS (2017). *Plan de gestion de données : fiche synthétique* (DOI : 10.13143/CGV4- 0K53). En ligne : https://doranum.fr/plan-%20gestion-donnees-dmp/fiche-synthetique

InIST-CNRS (s.d.). Site *DMP Opdidor*. https://dmp.opidor.fr

1.3. Sécurité informatique

ANSSI, Agence nationale de la sécurité des systèmes d'informations (2021). *Guide d'hygiène informatique – v.2*. 2017. En ligne : https://www.ssi.gouv.fr/guide/guide-dhygiene-informatique
Voir notamment :
https://www.ssi.gouv.fr/administration/bonnes-pratiques
https://www.ssi.gouv.fr/administration/precautions-elementaires/calculer-la-force-dun-mot-de-passe

SupDPO, Réseau des délégués à la protection des données de l'Enseignements supérieur et de la recherche et Réseau des responsables recherche (R3Sup) (16 jan. 2020). *Quinze recommandations aux chercheurs sur la protection des données dans le cadre de leurs activités de recherche* – v.1.
En ligne : https://reseau.supdpo.fr/wp-content/uploads/2020/03/SupDPO-Recommandations-chercheurs-1.2-1.pdf

2. Aspects juridiques de la réutilisabilité

2.1. Textes de référence

Code des relations entre le public et l'administration, livre III, titre II : La réutilisation des informations publiques, art. L321-1 et suivants.
En ligne : https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000031366350/LEGISCTA0000313 67685/#LEGISCTA000031367685

Code du patrimoine.
En ligne : https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006074236

Code de la propriété intellectuelle.
En ligne : https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT 000006069414

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (2016).
En ligne : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746

2.2. Guides juridiques pour la recherche

GINOUVES V., GRAS I. (dir.) (2018). *La Diffusion numérique des données en SHS. Guide des bonnes pratiques éthiques et juridiques.* PUP, coll. « Digitales ».

MAUREL, L.(s.d.) *S.I.Lex*. En ligne : https://scinfolex.com

GRAS, I., GINOUVES V. (dir.) (s.d.). Carnet de recherche *Questions d'éthique et droit en SHS*, Hypotheses : OpenEdition, En ligne : https://ethiquedroit.hypotheses.org

2.3. Sur les données publiques

CADA, CNIL, Etalab (2020). *Guide de la publication en ligne et de la réutilisation des données publiques (« Open Data »).* En ligne : https://www.cnil.fr/sites/default/files/atoms/files/guide-open-data.pdf

CNIL (2019). *Publication en ligne et réutilisation des données publiques (« open data »)*. En ligne : https://www.cnil.fr/fr/publication-en-ligne-et-reutilisation-des-donnees-publiques-open-data
Voir notamment : *Comment réutiliser les données diffusées ? :*
En ligne : https://www.cnil.fr/fr/comment-reutiliser-les-donnees-diffusees

MAUREL, L. (2018). « La Réutilisation des données de la recherche après la Loi pour une République numérique ». In *La Diffusion Numérique Des Données En SHS - Guide de Bonnes Pratiques Éthiques et Juridiques* (*op.cit*.). En ligne : https://hal.archives-ouvertes.fr/hal-01908766

MAUREL, L. (14 sept. 2020). « À qui appartiennent les données par Lionel Maurel », *Mate-SHS*, conférence.
En ligne : https://mate-shs.cnrs.fr/actions/tutomate/tuto25-propriete-donnees-lionel-maurel

RICARD, B. (dir.) (s.d.). Carnet de recherche *Droit(s) des archives*. OpenEdition-Hypotheses.
En ligne : https://siafdroit.hypotheses.org

2.4. Sur le droit d'auteur

Bureau de la propriété intellectuelle du Ministère de la culture (2021). *Fiches techniques sur les droits d'auteur et les droits voisins*.
En ligne : https://www.culture.gouv.fr/Thematiques/Propriete-litteraire-et-artistique/Fiches-techniques-sur-les-droits-d-auteur-et-les-droits-voisins2

DORD-CROUSLE S., GRELOU E., HUE E. et PIERROT D. (s.d.). *L'Édition numérique de corpus d'auteurs – aspects juridiques*, Consortium Cahier.
En ligne : https://cahier.hypotheses.org/corpus-auteurs-aspects-juridiques

LUCAS, A (dir.) (2018). *Guide du droit d'auteur.* 4ème édition. Université de Nantes, Institut de recherches en droit privé, MESRI. En ligne : http://media.sup-numerique.gouv.fr/file/Licences_et_droit_d_auteur/03/0/Guide_du_droit_d_auteur_4e_ed_2018_1006030.pdf

2.5. RGPD, données sensibles et recherche

DEPLANQUE, Catherine et al. (2018). *RGPD : Fiches pratiques à destination des chercheurs*.
En ligne : https://recherche.parisnanterre.fr/accueil/rgpd-fiches-pratiques-rgpd-recherche

InSHS (février 2021). *Les Sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte – Guide pour la recherche, version 2.*
En ligne : https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf

2.6. Licences de réutilisation

Creative commons (s.d.), « À propos des licences »
En ligne : https://creativecommons.org/licenses/?lang=fr-FR

ETALAB (2017). *Licence ouverte version 2.0*.
En ligne : https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf

3. Documentation des données

Les guides méthodologiques des consortiums Cahier et Corli :

CORLI (s. d.), *Bonnes pratiques*. Consortium Corli. En ligne : https://corli.huma-num.fr/bonnes-pratiques/
Voir en particulier : « Bonnes pratiques pour la constitution de corpus », https://corli.huma-num.fr/bonnes-pratiques-pour-la-constitution-de-corpus

CAHIER (s. d.), *Guides*. Consortium Cahier. En ligne : https://cahier.hypotheses.org/guides
Voir en particulier : guide du groupe « Droits et questions juridiques » (*op. cit.*)
« La publication des éditions de textes : Informations et recommandations », https://cahier.hypotheses.org/publication-editions-textes et le guide du groupe « Correspondance », https://cahier.hypotheses.org/guide-correspondance

BURNARD, L. (2018). *Comment maîtriser le tigre TEI*. Consortium Cahier.
En ligne : https://cahier.hypotheses.org/files/2018/08/ODD-diapos.pdf

CAMPS, J.-B. (2018). *Structuration des données et des documents : balisage XML*. Master.
En ligne : https://halshs.archives-ouvertes.fr/cel-01706530

ETALAB (2021). « Documenter les données », in *Préparer les données à l'ouverture et à la circulation*.
En ligne : https://guides.etalab.gouv.fr/qualite/documenter-les-donnees/#description-generale-du-jeu-de-donnees

URFIST Méditerrannée (2019). *Les principes FAIR*. Site *Doranum*.
En ligne : https ://doranum.fr/enjeux-benefices/principes-fair