Plan de Gestion de Données du Consortium CAHIER (PGD)

Fatiha Idmhand, Laurene L'Hermite

Plan de Gestion de Données du Consortium CAHIER

PGD V1: 30/10/2021, PGD Consortium CAHIER de Huma-Num

Ceci est le Plan de Gestion des Données du consortium CAHIER de l'Infrastructure Huma-Num. Il s'agit d'un plan de gestion des données de structure. De 2011 à 2021, CAHIER a travaillé à la numérisation et à la construction de données sur des corpus d'auteurs. Ce plan de gestion des données offre une vue générale des actions menées par le consortium sur les données au bout de dix ans.

CAHIER a donc rédigé deux types de plans de gestion des données : un Plan dit « de structure » et un plan par projet volontaire.

Les plans proposés par CAHIER s'inspirent des recommandations de l'Agence Nationale de la Recherche et du guide publié en 2016 par le programme Horizon Europe[1][1]. Les modèles mis en ligne par la plateforme DMP OPIDoR (https://opidor.fr/), l'outil d'aide à la création en ligne de plans de gestion de données (Data Management Plan ou DMP), ont été étudiés et adaptés aux besoins de la communauté scientifique du Consortium CAHIER. Le modèle de PGD proposé par CAHIER sera mis en ligne sur Opidor.

Auteurs du plan de gestion des données :

IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182, Université de Poitiers, Institut des textes et Manuscrits Modernes (ITEM, UMR8132), Paris, France

Rôle dans le consortium CAHIER : Coordinatrice adjointe

L'HERMITE, Laurène, Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France

Rôle dans le consortium CAHIER : Chargée de PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD Consortium CAHIER de Huma-Num
Une seule version de ce PGD est actuellement prévue

[1] [1] H2020 Programme « Guidelines on FAIR Data Management in Horizon 2020 », https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10

SOMMAIRE

Table des matières

PARTIE I

PGD V1: 30/10/2021, PGD du Consortium CAHIER de Huma-Num

PLAN DE GESTION DES DONNÉES DE STRUCTURE

LE CONSORTIUM CAHIER

1) Informations sur le plan de gestion de données (PGD)

- ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

- ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182, Université de Poitiers, Institut des textes et Manuscrits Modernes (ITEM, UMR8132), Paris, France
Rôle dans le consortium CAHIER : Coordinatrice adjointe

L'HERMITE, Laurène, Université de La Rochelle, Centre de recherches en histoire internationale et atlantique (EA1163), La Rochelle, France
Rôle dans le consortium CAHIER : Chargée de PGD

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD Consortium CAHIER de Huma-Num
Une seule version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...).
Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France

Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

Consortium CAHIER"Corpus d'auteurs pour les humanités : informatisation, édition, recherche" de l'infrastructure Huma-Num (2011-2021)

Responsable du projet (principal researcher) et unité de rattachement

- Responsable du Consortium de 2011 à 2015 :

DEMONET, Marie-Luce, Université de Tours, Centre d'études Supérieures de la Renaissance (UMR7323)

Rôle dans le consortium CAHIER : Coordinatrice de 2011 à 2015

- Co-responsables du Consortium de 2015 à 2021 :

LEBARBE, Thomas, Université de Grenoble-Alpes, (UMR5316), Grenoble, France

Rôle dans le consortium CAHIER : Coordinateur de 2015-2021

IDMHAND, Fatiha, IdHAL : fatiha-idmhand ; ORCID : 0000-0001-7135-9182, Université de Poitiers, Institut des textes et Manuscrits Modernes (ITEM, UMR8132), Paris, France

Rôle dans le consortium CAHIER : Coordinatrice adjointe de 2015-2021

Rôle dans le consortium CAHIER : Coordinatrice du 09/09/2021-31/12/20212021

Financeur(s) du projet et type de financement

Infrastructure Huma-Num, financement de consortiums
CNRS UAR 3598 Huma-Num
Bâtiment de recherche Nord, 14, cours des humanités, 93322 Aubervilliers cedex

Référence de la convention de financement

Infrastructure Huma-Num, financement de consortiums

Institution / organisme / unité porteuses du projet

Centre national de la recherche scientifique, CNRS

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

Maison des Sciences de l'Homme Val de Loire (MSH VdL) - USR CNRS 3501 -
33, Allée Ferdinand de Lesseps , 37204 TOURS CEDEX 03

Descriptif et objectif(s) du projet

CAHIER est un consortium interdisciplinaire de projets numériques menés principalement dans les domaines des ”corpus d’auteurs”. Constitué en fédération en 2011 dans le cadre de l’infrastructure”CORPUS”il a ensuite été intégré à la TGIR Huma-Num.
L’ensemble des projets sur corpus d’auteurs membres du consortium ont une activité éditoriale numérique ou double support (sur papier et en ligne). Ces éditions comportent différents degrés de description et d’analyse allant de l’édition du texte brut avec apparat critique minimal jusqu’à l’édition critique complète, l’édition génétique associant aux textes les images des documents originaux (manuscrits ou imprimés), en passant par des données encyclopédiques (dictionnaires, pièces d’archives, illustrations, bases de données). Ce consortium se définit par rapport à l’existence d’une œuvre (incluant les documents préparatoires) ou de plusieurs œuvres identifiées, dont la cohérence mérite d’être soulignée, publiée et outillée pour donner lieu à de nouvelles recherches. Il a pour but :

- d’augmenter l’acquisition de données de qualité (image et texte) tout en tenant compte des limites de taille
- de proposer et partager des normes de transcription en suivant les objectifs éditoriaux clairement énoncés
- de permettre l’indexation des corpus et des images
- d’offrir des données et des métadonnées compatibles avec les standards internationaux qui permettent l’exploitation des données (catalogage, archivage, identification, protection des données, moissonnage, etc.)
- d’adapter les solutions juridiques et les modèles de convention de partenariat (établir un dialogue efficace et collectif avec les éditeurs) en lien avec l’évolution des pratiques numériques
- d’offrir les moyens d’évoluer vers le web sémantique, la visualisation, les entrepôts de données, les modes de représentation d’ensembles documentaires, et vers l’annotation collaborative
- de favoriser l’appropriation des outils numériques en organisant des formations et des ateliers.

Le consortium permet de tester, voire d’élaborer, des outils prototypes (édition, traitement des données, etc.) dans le domaine des humanités numériques. Il favorise, en contribuant à leur financement, la participation aux ateliers et congrès internationaux et aux manifestations organisées par les structures européennes.

Dates et durée

Date de début de financement et de début des travaux : 2011
Date de fin de financement et de fin des travaux : 2021

Mots clés du projet

Corpus, Auteurs, Humanités, Informatisation, Édition, Recherches

Publications (articles, pré-proposition, site web, …)

Site web du consortium : https://cahier.hypotheses.org

Listes des articles publiés par le Consortium : https://halshs.archives-ouvertes.fr/search/index/q/*/structId_i/545625/
Autres livrables (guides, recommandations, etc.) : https://cahier.hypotheses.org/guides

3) Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doivent être décrits. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

Consortium CAHIER

Corpus

CAHIER est un consortium interdisciplinaire qui associe des projets numériques du domaine des "corpus d'auteurs». Qu'ils relèvent de la littérature, de la philosophie ou d'une thématique liée à une école ou à une pratique, les corpus étudiés par ce consortium sont, le plus souvent, associés à une activité éditoriale numérique, le plus souvent sur double support (papier et en ligne).

Période couverte par le corpus, auteur(s) concerné(s)

Le consortium se définit par rapport à l'existence d'une œuvre (incluant les documents préparatoires) ou de plusieurs œuvres identifiées, dont la cohérence mérite d'être soulignée, publiée et outillée pour donner lieu à de nouvelles recherches. Les corpus du consortium couvrent toutes les périodes depuis l'Antiquité, tous les genres littéraires et toutes les spécialités des sciences des textes.

Organisation du corpus

Les corpus numériques sont créés par les projets membres, ils comportent différents degrés de description et d'analyse allant de l'édition du texte brut avec apparat critique minimal, du corpus d'images accompagné de métadonnées jusqu'à l'édition critique complète ou l'édition génétique.

Mode de collecte et origine des données

Les projets numériques membres du consortium CAHIER collectent et constituent leurs données numériques de la façon suivante:

- lorsqu'il s'agit d'images: numérisations des sources primaires par le projet ou achat d'images numériques auprès d'opérateurs ayant déjà numérisé ces sources
- lorsqu'il s'agit des métadonnées : production automatique de métadonnées lors de la numérisation, saisie manuelle de métadonnées destinées à enrichir les données, saisie manuelle de métadonnées enrichissant l'édition (lors d'édition XML-TEI)
- lorsqu'il s'agit de données disponibles en ligne : réutilisation de données publiées, collecte semi-automatisée de données disponibles en ligne, traitement semi-automatisé des données, réutilisation

Les données constituées, collectées et étudiées par le consortium sont décrites selon les standards et normes actuelles:

- les archives sont décrites en XML EAD ou en Dublin Core selon les normes et recommandations des domaines
- les images sont nommées par les institutions qui hébergent les documents numérisés, lorsqu'il s'agit d'équipe de recherches, elles respectent les recommandations *du domaine: https://doranum.fr/stockage-archivage/comment-nommer-fichiers/*
- les éditions numériques sont produites et les données décrites en XML TEI

Etat du corpus numérique

Le consortium CAHIER a fondamentalement produit trois "types" de publications numériques : des "archives éditorialisées", des PGD V1: 30/10/2021, PGD du Consortium CAHIER de Huma-Num (2011-2021) des "éditions de lecture" et des "éditions enrichies"[2].
A la différence du guide des recommandations d'autres consortiums comme celui de la TEI et des *Best Practices for TEI in Libraries*, les publications numériques prises en compte par le consortium CAHIER ne sont pas uniquement codées en XML TEI, d'autres options éditoriales et techniques, susceptibles d'être utilisées par les chercheurs, ont été prises en compte, comme par exemple la publication via un CMS. Toutefois, pour garantir l'accessibilité et l'évaluation des publications, CAHIER a fortement recommandé d'utiliser des standards de description de données partagés par les communautés académiques internationales.

- le métalangage de description de documents XML permet la structuration poussée des contenus manipulés ainsi que leur exploitation dans des contextes variés. Ainsi, le consortium CAHIER a très largement recommandé, encouragé et utilisé le vocabulaire de référence TEI pour l'encodage de documents textuels ou nativement numériques et suggéré l'EAD pour l'encodage des objets de recherche archivistiques ;
- dans le cadre d'un projet de publication ayant vocation à être largement diffusé dans un contexte d'Open Access, l'usage de vocabulaires standardisés pour décrire les documents ca été particulièrement encouragé en lieu et place de la création de vocabulaires ad hoc potentiellement équivoques et abscons, le Dublin Core a fait partie des préconisations de CAHIER.

[2] Voir à ce propos : Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations (V1-Novembre 2018), https://halshs.archives-ouvertes.fr/halshs-01932519

Types de données:

Les données numériques produites par les projets membres du Consortium CAHIER sont de quatre types:

- images : JPEG, PDF et TIFF
- textes : txt, xml tei, JPEG, TIFF et PDF
- audio et son : mp3, mp4
- vidéos : avi, flv

Volumétrie

La production des données des 65 projets membres a été estimée à :

- 50 sites webs
- 30 URL permettant d'accéder aux téléchargements direct des fichiers
- 327450 fichiers annoncés comme disponibles mais 13201 fichiers réellement disponibles, pas d'informations précises sur la présence ou non d'océrisations
- 500000 images sources potentiellement disponibles mais moins de 35000 réellement disponibles, impossible d'estimer le poids en Go à ce stade

Modifications effectuées sur les données, versions, …

CAHIER n'a réalisé aucun traitement sur les fichiers produits par ses membres. Les projets membres sont les seuls à modifier leurs fichiers.
CAHIER a largement encouragé la FAIRisation des données et le dépôt de celles-ci dans l'entrepôt fourni par l'infrastructure Huma-Num : Nakala.
Les membres de CAHIER ont également utilisé Zenodo et Ortolang pour FAIRiser leurs données

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

A ce stade, CAHIER n'a réalisé aucun enrichissement des données produites par les projets membres.

Métadonnées, créées et standards et formats utilisés

Les métadonnées produites par les projets membres du Consortium sont destinées à décrire les ressources produites, les classer, organiser et à caractériser leur contenu. CAHIER a produit, de façon quasi-systématique, au moins quatre types de métadonnées pour ses publications numériques :

- des métadonnées descriptives permettant l'identification non ambiguë de la source (analogique et/ou numérique) ;
- des métadonnées administratives apportant des informations sur les caractéristiques des fichiers, les droits d'accès et d'usage, et sur le processus de création des données ;
- des métadonnées structurelles expliquant la composition ou l'organisation de la ressource : pages, chapitres, table des matières ou autres éléments constitutifs. Ces métadonnées facilitent la caractérisation, la navigation, la présentation et la compréhension de la structure des sources ;
- des métadonnées techniques précisant les caractéristiques techniques des données, les logiciels utilisés pour leur production et manipulation, leurs versions ;

- et des annotations permettant d'analyser et d'interpréter la source ou métadonnées d'enrichissement (balisage) faites au fil du texte, au moyen d'un jeu réfléchi et prédéterminé d'étiquettes et d'attributs.

Les métadonnées descriptives, administratives et techniques

CAHIER a recommandé qu'un certain nombre de champs soient complétés et notamment les champs mentionnés par la norme Dublin Core non-étendu. Ont été préconisés, en tant que minimum souhaitable, les champs qui apportent des informations sur le texte (titre, éditeur, date), la propriété intellectuelle (auteur, droits), l'instanciation, la gestion, la ressource (formats, dimensions), son contenu (type, mots-clés) et les modalités de préservation des documents. Dans une édition XML/ TEI, ces informations se retrouvent dans la structure minimale de l'en-tête (header TEI).

Lors de la saisie des métadonnées, CAHIER a incité les projets membres à porter une attention soutenue à la normalisation de la présentation de celles-ci, ce qui implique, par exemple, le respect des recommandations internationales pour la saisie des dates (AAAA-MM-JJ), des noms de lieux (PAYS, Ville), des noms de personnes (NOM, Prénom), etc. Le cas échéant, il pourra être utile de recourir à des thésaurus pour faciliter l'exploitation postérieure des données.

Dans le cas d'une publication de type "Édition enrichie" (type 3 du guide publié en 2018[3]), on présentera obligatoirement un jeu de métadonnées plus étendu en apportant toutes les informations nécessaires à la description du témoin de départ et à la caractérisation de la publication effectuée. Dans ce cas, un standard reconnu d'encodage des métadonnées, tel que METS, MIX, UNIMARC, XML-EAD, Dublin Core simple/étendu est recommandé, et lorsqu'il s'agit d'une édition XML/TEI, plusieurs éléments et sections du header permettent d'atteindre un très haut niveau de précision et de finesse dans la description de la source de départ et de l'édition produite. CAHIER a préconisé ainsi, qu'en plus de la section obligatoire <fileDesc>, que les sections <encodingDesc>, <profileDesc> et <revisionDesc> soient renseignées.

Les métadonnées structurelles et l'annotation sémantique

Dans une édition papier, les notes constituent traditionnellement l'espace privilégié dédié à l'apport des informations scientifiques. Tout en permettant d'insérer des notes, l'édition électronique dispose de systèmes plus élaborés pour enrichir le texte que CAHIER a recommandé d'utiliser. La présence d'enrichissements sémantiques, à l'aide de balises par exemple, constitue un élément discriminant entre les différents types de publications numériques.

L'enrichissement de la publication peut être réalisé à l'aide de multiples technologies et outils (traitement ou éditeur de texte, éditeur xml ou html, etc.) mais selon les projets de publication numériques, CAHIER a encouragé ses projets membres, selon les cas, à :

- coder, dans le cas des "éditions de lecture", les grandes sections du texte (, dans la TEI), les titres (lorsqu'ils existent, <head> dans la TEI) et les paragraphes (à l'aide des balises dédiées) de façon générale dans le cas ; également les actes, scènes et tours de parole pour le théâtre ; les strophes et les vers pour la poésie ; les et les paragraphes, les destinataires, les expéditeurs, les lieux de rédaction et d'expédition et les dates pour les correspondances ; les sections et les articles, les

[3] Cf. : https://halshs.archives-ouvertes.fr/halshs-01932519

rédacteurs, les dates de parution des textes pour les journaux, revues et gazettes ; et et de séparer le mot vedette du texte de l'entrée à proprement parler pour les entrées des dictionnaires et des encyclopédies.

- coder, dans le cas des "éditions enrichies", les divisions et les différents éléments de structure (scène et actes ; chapitres ; vers ou groupes de vers ; articles, etc.) avec une granularité aussi fine que possible. Dans le cas du théâtre, on pourra trouver des attributs (comme @who) permettant d'identifier de façon non ambiguë le locuteur de chaque réplique, et des didascalies clairement typées ; dans le cas de la poésie, des éléments annotant la rime et le rythme ; dans le cas des dictionnaires, une annotation de l'information grammaticale, étymologique et d'usage (si possible), l'identification des sources des citations et leur équipement avec des liens hypertexte pointant vers des éditions extérieures, etc.

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

Le Consortium CAHIER a laissé chaque projet membre utiliser les vocabulaires contrôlés propres au champ du corpus d'auteurs ou du projet de recherche. Le plus souvent, les membres ont utilisé les vocabulaires contrôlés de thésaurus tels que le RAMEAU (référentiel d'autorité - matière en usage dans le milieu de la documentation et des bibliothèques d'enseignement supérieur et par la BnF).

A partir de 2015, et pour combler un manque concernant le vocabulaire permettant de décrire les concepts du domaine des textes littéraires et plus particulièrement les "genres" ou "types" de textes, CAHIER à travaillé à l'élaboration d'un référentiel pour les genres textuels : le thésaurus "Typologie textuelle" (https://opentheso.huma-num.fr/opentheso/ : ouvrir "Typologie 43"). CAHIER a utilisé le logiciel opensource Opentheso pour l'élaborer. Les thésaurus Opentheso sont édités et consultés en ligne et peuvent être importés et exportés sous plusieurs formats, et notamment SKOS[4] (SKOS Reference 2009). Un identifiant pérenne de type Handle ou Ark peut être attribué à chaque concept.[5]

Dans ce thésaurus, 365 concepts ont été décrits et tous ont été pourvus d'un identifiant unique de type DOI. CAHIER recommande aux projets membres d'utiliser les URL de ces identifiants, de les insérer dans les métadonnées pour décrire le genre textuel ou type de leurs documents.

[4] Langage de représentation de schémas de concepts mis au point par le W3C. Le SKOS permet de gérer des modèles sémantiques relationnels (type thésaurus, index d'autorités matière, taxonomies, folksonomies, …) de façon simple, dans la perspective du web sémantique.

[5] La documentation complète d'Opentheso est disponible en ligne sur le site https://opentheso.hypotheses.org.

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être important de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

Les projets membres du Consortium ont fait appel à leurs institutions pour organiser le stockage de leurs données pendant la durée de leurs projets.
Dans quelques cas épars, ces institutions ont également la capacité de fournir des services de stockage à long terme et des identifiants pérennes aux données de façon à faciliter le stockage et l'accessibilité de celles-ci.
De façon générale, CAHIER a préconisé l'utilisation des services d'Human-Num pour assurer :

- la préservation des données pendant la durée des projets (le service Sharedocs a été largement utilisé)
- le stockage des données à long terme via l'entrepôt de stockage Nakala permettant de rendre les données des projets FAIR.

En 2021, les préconisations du Consortium en vue d'accompagner et de simplifier la FAIRisation des données des projets grâce à l'outil Nakala[6] ont essentiellement concerné les métadonnées descriptives. CAHIER a recommandé le respect des métadonnées suivantes et indiqué aux projets les champs attendus par Nakala :

Champ requis par Nakala	Données décrites en Dublin Core par les projets / Données attendues par Nakala en DCterms	Données décrites en TEI par les projets / Données attendues par Nakala en DCterms
Titre	dc:title / dcterms:title	teiHeader/fileDesc/titleStmt/title @type="main" / dcterms:title

[6] Voir le guide FAIR du Consortium CAHIER : https://halshs.archives-ouvertes.fr/halshs-02889777

Auteur	dc:creator / dcterms:creator	teiHeader/fileDesc/titleStmt/author / dcterms:creator
Éditeur scientifique	dc:contributor / dcterms:contributor	teiHeader/fileDesc/titleStmt/editor / dcterms:contributor
Éditeur	dc:publisher / dcterms:publisher	teiHeader/fileDesc/publicationStmt/publisher / dcterms:publisher
Date de publication du fichier électronique	dc:issued / dcterms :dateIssued Ou dcterms:date available (date à laquelle la ressource est devenue ou deviendra disponible.)	teiHeader/fileDesc/publicationStmt/date / dcterms :dateIssued Ou teiHeader/fileDesc/publicationStmt/availability/licence/date (si présent) / Ou dcterms:date available
Date	dc:date / dcterms:dateCreated Date de création du document source (appelée date première dans le catalogue OAI du consortium CAHIER)	teiHeader/profileDesc/creation/date @when="1857" / dcterms:dateCreated Ou teiHeader/profileDesc/creation/date @notBefore="1700" @notAfter="1750" / dcterms:dateCreated
Informations sur les droits d'utilisation	dc:rights / dcterms:rights	teiHeader/fileDesc/sourceDesc/bibl/note[@type='settlement'] / dcterms:rights ou teiHeader/filedesc/sourceDesc/biblStruct/note[@type='settlement'] / dcterms:rights ou, pour les manuscrits, teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/ settlement / dcterms:rights
Identifiant	dc:identifier / dcterms:identifier	teiHeader/fileDesc/publicationStmt/idno[@type="URL"] / dcterms:identifier Dans un second temps on ajoutera teiHeader/fileDesc/publicationStmt/idno @type="DOI" / dcterms:identifier
URI de licence	dc:rights / dcterms:license	teiHeader/fileDesc/publicationStmt/availability/licence[@target="URI de la licence"] / dcterms:license

Référence (bibliographique) du document d'origine	dc:source / dcterms:source	teiHeader/fileDesc/sourceDesc/bibl/note[@type='identifier'] / dcterms:source ou pour une bibliographie plus détaillée : teiHeader/fileDesc/sourceDesc/biblStruct/note[@type='identifier'] / dcterms:source Pour les manuscrits teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/[différents champs pertinents, don't idno] / dcterms:source
Langue	dc:language / dcterms:language	teiHeader/profileDesc/langUsage/language @ident="fr" / dcterms:language
Résumé	dc:description / dcterms:description	teiHeader/profileDesc/abstract / dcterms:description
Mots-clés	dc:subject / dcterms:subject	teiHeader/profileDesc/textClass/keywords/term @type="subject" / dcterms:subject

Accès, partage et limites des données

Les données produites par les projets membres du Consortium CAHIER sont d'emblée entièrement accessibles sans restriction. Dès sa fondation en 2011, CAHIER a fait de l'ouverture des données une priorité. Lorsque certaines données posent des contraintes aux ou des modalités spécifiques de diffusion, elles sont publiées sous embargo, limité dans le temps ou sous une Licence qui limite le partage des et explicite clairement les conditions et raisons particulières de consultation ou de réutilisation (nécessité d'un logiciel par exemple, d'un mot de passe, etc.).

La majorité des données diffusées par CAHIER sont rendues publiques en accès libre, à partir de textes et d'images intégralement libres de droits ou avec autorisation des ayants-droits.

En revanche, les métadonnées sont rendues disponibles sous la forme d'apparat critique et de notes explicatives sans restrictions, sous réserve du respect de la propriété intellectuelle de leurs auteurs et des documents auxquels elles se réfèrent.

5) Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

Le consortium CAHIER ne dispose pas de personne ressource dédiée à la gestion des données. Chaque projet membre gère ses données dans le cadre de son laboratoire, de sa MSH ou de son université.
En revanche, CAHIER assure la mission de relais avec l'infrastructure à l'étape du stockage pérenne des données. Les données déposées dans Nakala sont gérées par l'infrastructure Huma-Num qui assure le stockage pérenne de celles-ci, leur délivre des identifiants pérennes (DOI) et propose des services permettant de les rendre interopérables et moissonnables par l'intermédiaire du protocole OAI-PMH.
Une fois sur Nakala, le responsable de la gestion des données est Huma-Num.
Avant ce dépôt, le :

- responsable de la qualité des données (traitement, anonymisation, format, nettoyage,…) est : le responsable du projet membre
- responsable de la collecte des données est : le responsable du projet membre
- responsable du stockage et de la sécurité des données (s'il ne s'agit pas d'un service d'Huma-Num) : le responsable du projet membre

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6) Archivage des données

- ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

CAHIER utilise et recommande l'entrepôt de données Nakala

Durée de conservation des données

A ce stade, CAHIER recommande le stockage illimité dans le temps dans l'entrepôt de données Nakala.

Volume des données à conserver

L'ensemble des données déposées sur Nakala est voué à être conservé sur le long terme, il représente actuellement, et à ce stade des travaux du Consortium plus de 60Go.

Coûts alloués à la conservation

Les coûts d'archivage sont assumés par Huma-Num via Nakala. L'économie est estimée à près de ~7000€ / an pour chaque projet membre

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

Les outils utilisés par les membres du Consortium pour produire leurs données sont normalement libres, ouverts et pérennes. Aucun logiciel spécifique ne devrait être nécessaire pour accéder aux données.

7) Partage des données à l'issue du projet

- ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s'il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR : les données doivent être Trouvables (Findables), Accessibles, Interopérables et Réutilisables. Pour être réutilisables, les données doivent être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l'accompagnement d'une description et de documentation, par des formats ouverts et non propriétaires et par une disponibilité garantie par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

Les données produites par le Consortium CAHIER concernent des corpus issus du domaine des sciences des textes littéraires, de tous les genres textuels de ce domaine et couvrent un empan large allant de l'Antiquité à nos jours. Le potentiel de réutilisation des données produites est élevé, bien que ces corpus numériques s'adressent, dans leur grande majorité, à une communauté experte et spécialiste du champ disciplinaire, de la période et/ou de l'auteur. Les données publiées par CAHIER visent essentiellement à documenter la recherche internationale sur les corpus concernés, la publication de ces données apporte des informations permettant d'éclaircir des zones d'ombre..
Si ces données s'adressent avant tout à un public de chercheurs en sciences humaines et sociales, elles peuvent néanmoins intéresser un public d'amateurs.

Eléments d'accompagnement qui permettent la réutilisation des données.

Comme indiqué précédemment, le consortium CAHIER a veillé à accompagner et recommander un certain nombre de bonnes pratiques facilitant la réutilisation des données produites par le consortium:

- les données produites respectent les standards et bonnes pratiques de numérisation des domaines ;
- les jeux de données sont accompagnés d'au moins quatre à cinq types de métadonnées afin d'en faciliter la contextualisation, la compréhension et la réutilisation à long terme ;
- les données produites sont documentées ;
- les données produites sont ouvertes, libres, réutilisables ;
- les données sont stockées dans un entrepôt ouvert (Nakala notamment) et sûr.

Publications sur les données pour en améliorer l'exposition

Les membres du consortium CAHIER ont publié de nombreux travaux en vue de disséminer leurs résultats scientifiques et de promouvoir et donner à connaître les données produites. Certains de ces travaux peuvent être lus et consultés sur HAL : https://halshs.archives-ouvertes.fr/search/index/q/*/structId_i/545625/ et sur les sites web des projets.

Conditions de réutilisation : licences et contrats pour l'ensemble du projet

Chaque projet membre a assorti ses publications numériques et données de la licence la plus adaptée en fonction de son projet.
Toutefois, CAHIER a recommandé l'usage de Licence Creative Commons, et tout particulièrement, des licences suivantes afin d'encourager la réutilisation et la réexploitation des données :

Conditions et modes d'accès	**Licence ou contrat**	**Embargo / confidentialité**
Lorsque l'accès est totalement libre et la donnée réutilisable et modifiable sans restriction (CC) La citation de l'origine de la donnée ou source est obligatoire (BY)	Licence CC - BY	Pas d'embargo
Lorsque l'accès est totalement libre et la donnée réutilisable (CC) mais sous les conditions suivantes : - pas d'utilisation commerciale = NC - pas de modification de la source (ND) - obligation de rediffuser selon les mêmes conditions (SA) La citation de l'origine de la donnée ou source est obligatoire (BY)	Licence CC - BY - NC Licence CC - BY - NC - ND ou Licence CC - BY - NC - SA	Pas d'embargo mais si embargo il y a, obligation de préciser la durée des restrictions

PARTIE II

PGD V1: 30/10/2021, PGD du Consortium CAHIER de Huma-Num

MODÈLE ET RECOMMANDATIONS POUR LA RÉDACTION D'UN PLAN DE GESTION DES DONNÉES SUR DES CORPUS D'AUTEURS

1) Plan de gestion de données (PGD) du projet XXXXXXXX

- ***Présentation de la section***

Cette section décrit le PGD: elle présente l'auteur du PGD, les relecteurs du PGD, les autres intervenants assurant la gestion du PGD et, le cas échéant,ses mises à jour.

- ***Recommandations :***

Il est utile de désigner un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Il est recommandé d'associer ce responsable à son identifiant ORCID, IdRef, ISNI, IdHal et de nommer l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Le PGD évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet et d'indiquer les versions du PGD dans leur ordre antéchronologique en commençant par l'actuelle.

Auteur du plan de gestion des données :

NOM, Prénom, IdHAL : xxxxxxxxx ; ORCID : xxxxxxxxx, Université de xxxxxxxxx, xxxxxxxxx (sigle, EA ou UMR n°xxxx), xxxxxxxxx, France
Rôle dans le projet : xxxxxxxxx

NOM, Prénom, IdHAL : xxxxxxxxx ; ORCID : xxxxxxxxx, Université de xxxxxxxxx, xxxxxxxxx (sigle, EA ou UMR n°xxxx), xxxxxxxxx, France
Rôle dans le projet : xxxxxxxxx

Version du plan de gestion des données :

PGD V1: 30/10/2021, PGD projet xxxxxxxxx
xxxxxxxxx version de ce PGD est actuellement prévue

2) Présentation du projet et responsabilités

- ***Présentation de la section***

Cette section décrit le projet ou le corpus sur lequel porte le PGD. Elle décrit le projet, ses objectifs, participants, etc. Ici, nous décrivons le Consortium CAHIER mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Si le nom du projet est un acronyme, indiquez également la version développée.
Exemple : Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")
Identifier la/le responsable scientifique du projet : Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...).
Si possible indiquez des données de contact (courriel, téléphone professionnel)
Exemple : Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Nom du projet

xxxxxxxxx

Responsable du projet (principal researcher) et unité de rattachement

NOM, Prénom, IdHAL : xxxxxxxxx ; ORCID : xxxxxxxxx, Université de xxxxxxxxx, xxxxxxxxx (sigle, EA ou UMR n°xxxx), xxxxxxxxx, France
Rôle dans le projet : xxxxxxxxx

Financeur(s) du projet et type de financement

xxxxxxxxx

Référence de la convention de financement

xxxxxxxxx

Institution / organisme / unité porteuses du projet

xxxxxxxxx

Partenaires (identifier les organismes partenaires, ressources et co-financeurs du projet)

xxxxxxxxx

Descriptif et objectif(s) du projet

xxxxxxxxx

Dates et durée

Date de début de financement et de début des travaux : xxxxxxxxx
Date de fin de financement et de fin des travaux : xxxxxxxxx

Mots clés du projet

xxxxxxxxx

Publications (articles, pré-proposition, site web, …)

Site web du projet : xxxxxxxxx
Listes des articles publiés par le projet : xxxxxxxxx
Autres livrables (guides, recommandations, etc.) : xxxxxxxxx

3) Présentation et description du corpus

- ***Présentation de la section***

Cette section décrit le corpus et ses données. Elle décrit de façon plus précise les données du projet, les méthodes appliquées pour les collecter, etc. Ici, nous décrivons les données du Consortium CAHIER dans leur ensemble mais vous trouverez en annexe des exemples plus précis basés sur des projets membres de CAHIER

- ***Recommandations :***

Il s'agira de préciser le mode de collecte et l'origine des données, les centres d'archives, bibliothèques ou centres d'études hébergeant les données y compris si les données procèdent d'un moissonnage de ressources en ligne. L'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers doit être décrite. De même que la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état, etc. Pour que les données soient réutilisables sur le long terme, les formats doivent être ouverts et non propriétaires et les données stockées dans des entrepôts accessibles.

Nom du projet

xxxxxxxxx

Présenter et décrivez le corpus

xxxxxxxxx

Période couverte par le corpus, auteur(s) concerné(s)

xxxxxxxxx

Organisation du corpus

xxxxxxxxx

Mode de collecte et origine des données

xxxxxxxxx

Etat du corpus numérique

xxxxxxxxx

Types de données:

xxxxxxxxx

Volumétrie

xxxxxxxxx

Modifications effectuées sur les données, versions, …

xxxxxxxxx

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

xxxxxxxxx

Métadonnées, créées et standards et formats utilisés

xxxxxxxxx

Les métadonnées descriptives, administratives et techniques

xxxxxxxxx

Les métadonnées structurelles et l'annotation sémantique

xxxxxxxxx

Référentiels d'indexation utilisés (vocabulaires contrôlés - thésaurus ou ontologies disciplinaires - et/ou indexation libre)

xxxxxxxxx

4) Modalités de partage, de sauvegarde et de protection des données. Volumétrie des données stockées et espaces choisis.

- ***Présentation de la section***

Cette section décrit la documentation produite au cours projet. Il s'agit d'une documentation autre que numérique (sur support papier par exemple). Si elle existe, il est important de la décrire. Cette section décrit également les lieux et infrastructures de stockage des données pendant le projet.

- ***Recommandations :***

Il s'agira de préciser ici le matériel physique et les lieux de stockage des données. Idéalement, il faudrait stocker les données dans au moins 2 endroits, éviter le stockage externe et privilégier les outils mis à disposition par l'institution. Pour cela, il peut être nécessaire de savoir quel est le volume approximatif des données à sauvegarder, l'espace de stockage nécessaire, la périodicité des sauvegardes, le nom et la nature du service fourni par l'institution, etc.On peut également indiquer les procédures de sauvegarde mises en place (fréquence des sauvegardes, automatisée ou non ?), les personnes en charge de la protection de ces données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…

xxxxxxxxx

Accès, partage et limites des données

xxxxxxxxx

5) Responsabilités et ressources pour la gestion des données

- ***Présentation de la section***

Cette section décrit, identifie, présente et nomme les responsables de la gestion des données.

- ***Recommandations :***

Afin de respecter les principes FAIR, CAHIER recommande le dépôt de celles-ci dans l'entrepôt Nakala (https://www.nakala.fr/). Ce service de dépôt et de stockage des données est proposé par la TGIR HumaNum pour les SHS. Il assure la gestion pérenne et sûre des données. Utiliser Nakala n'empêche pas de recourir à un second dépôt sur un autre entrepôt ou sur une plateforme institutionnelle.

xxxxxxxxx

Évaluation des coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Dans le cadre du Consortium CAHIER, les moyens assumés par l'infrastructure Huma-Num ont concerné les tâches suivantes:

- mise à disposition de moyens matériels tels que des serveurs, machines virtuelles, logiciels dédiés et licences supplémentaires dont les coûts et abonnements ne sont

pas supportés par les projets, soit une économie estimée à ~5000€ / an pour chaque projet membre
- mise à disposition de moyens humains (ETP) pour des tâches spécifiques relevant à la fois de la gestion des moyens matériels (serveurs, machines, etc.), du stockage des données et des actions de formation, soit une économie estimée à plus de ~50000€ / an pour chaque projet membre

6) Archivage des données

- ***Présentation de la section***

Cette section décrit les données à conserver à court, moyen et long terme, les éventuelles données à détruire ou à laisser sous embargo et indique la durée de cette restriction.

- ***Recommandations :***

A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire, elles pourront être destructibles pour des raisons de légalité ou de confidentialité.

Plateforme pour l'archivage pérenne des données

xxxxxxxxx

Durée de conservation des données

xxxxxxxxx

Volume des données à conserver

xxxxxxxxx

Coûts alloués à la conservation

xxxxxxxxx

Outils, méthodes, procédures nécessaires pour accéder à ces données archivées et les réutiliser

xxxxxxxxx

7) Partage des données à l’issue du projet

- ***Présentation de la section***

Cette section décrit la politique de dissémination des données. Elle indique s’il existe des limites à la diffusion des données, comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public.

- ***Recommandations :***

Une bonne dissémination des données requiert, dans la mesure du possible, le respect des principes FAIR: les données doivent être trouvables (findables), accessibles, interopérables et réutilisables. Pour être réutilisables, les données doivent être faciles d’accès, identifiables et citables grâce à des identifiants uniques (DOI) et leur usage facilité par l’accompagnement d’une description et de documentations, par des formats ouverts et non propriétaires et par une disponibilité facilitée par un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

xxxxxxxxx

Eléments d’accompagnement qui permettent la réutilisation des données.

xxxxxxxxx

Publications sur les données destinées à en améliorer l’exposition

xxxxxxxxx

Conditions de réutilisation : licences et contrats pour l’ensemble du projet

xxxxxxxxx

Annexe

RECOMMANDATIONS DÉTAILLÉES DU CONSORTIUM CAHIER POUR L'ÉLABORATION DES PLANS DE GESTION DES DONNÉES DES PROJETS SUR CORPUS D’AUTEURS

1) Informations sur le plan de gestion de données

Responsabilités (rédacteur du PGD, relecteurs, autres intervenants assurant la gestion du PGD et ses mises à jour)

Recommandations :
Désignez un responsable du PGD qui sera la personne à contacter. Il n'est pas nécessairement le responsable scientifique du projet. Si possible, associez-le à son identifiant ORCID, IdRef, ISNI, IdHal. Nommez ensuite l'ensemble des personnes ayant contribué à la rédaction et à la relecture du PGD.
Modèle / exemple :
Rôle : Nom, Prénom, Identifiant, Institution, Laboratoire, Unité de rattachement, Ville, Pays

Versions du document, historique des mises à jour et nombre de versions prévues

Recommandations :
Un PGD est un document qui évolue au fur et à mesure de l'avancée du projet de recherche et de l'enrichissement des données. Afin de faciliter sa rédaction, il est conseillé d'en produire une première version au début du projet, qui sera modifiée éventuellement en cours de projet, ainsi qu'à la fin du projet. Indiquez les versions dans l'ordre antéchronologique en commençant par l'actuelle.
Pour H2020 par exemple il est demandé 3 versions du PGD dont la première doit être envoyée dans les 6 mois après le début du projet.
Modèle / exemple :
PGD V2 : 08/09/2021, P. Nom, P. Nom, P. Nom, …
PGD V1 : 15/06/2021, P. Nom, P. Nom,...
Trois versions du PGD sont prévues.

2) Présentation du projet et responsabilités

Nom du projet

Recommandations :
Si le nom du projet est un acronyme, indiquez également la version développée.
Modèle / exemple :
Antonomaz ("ANalyse auTOmatique et NumérisatiOn des MAZarinades")

Responsable du projet (principal researcher) et unité de rattachement

Recommandations :
Nom, Prénom, Institution, Laboratoire, Unité de rattachement, Ville, Pays. Mettre en lien son identifiant ORCID (ou ISNI, IdRef, IdHal, ...). Si possible indiquez des données de contact (courriel, téléphone professionnel)
Modèle / exemple :

Karine ABIVEN (https://orcid.org/0000-0001-9518-1040), Sens-Texte-Informatique-Histoire (STIH), EA 4509, Université Paris - Sorbonne, Paris IV, France

Financeur(s) du projet et type de financement

Référence de la convention de financement

Institution / organisme / unité porteuses du projet

Recommandations :
Précisez également si le projet s'inscrit dans une programmation scientifique financée et les axes scientifiques liés à cette programmation :

- *Axes scientifique d'un Labex*
- *Programme de financement d'un projet ANR, H2020*
- *Axe ou programme scientifique d'une structure de recherche liée au porteur ou à l'équipe projet...*

Organismes partenaires, ressources et co-financeurs du projet

Modèle / exemple :
Projet dans le cadre de l'IUF, en partenariat avec la Bibliothèque Mazarine, et cofinancé par l'OBVIL (SU), le DIM STCN, le consortium CORLI.

Descriptif et objectif(s) du projet

Recommandations :
Description de la nature du projet, ses objectifs et son déroulement. Il s'agit de comprendre le contexte et les types de données qui seront produites ou collectées au cours du projet.

Dates et durée

Recommandations :
Dates de début de financement ou date de début des travaux pour les projets sans financement.
Date de fin de financement ou durée prévue des travaux pour les projets sans financement.

Mots clés du projet

Recommandations :
Utilisez dans la mesure du possible des vocabulaires contrôlés, des thésaurus - préciser lesquels - et indiquer si les termes disposent d'identifiants types DOI, URI ou permaliens et les lier.

Publications (articles, pré-propositions, site web, …)

Recommandations :
Publications dans le cadre du projet et listées dans l'ordre antéchronologique.
Modèle / exemple :

En 2020 : Mise en ligne du Thésaurus des poissons et créatures aquatiques. En 2021 il comprend 2290 entrées.
En 2020 : Mise en ligne de la Bibliothèque Ichtya : https://ichtya.unicaen.fr/lab/bibliotheque/

3) Présentation et description du corpus

Présentation et description du corpus

Mode de constitution du corpus, collecte et origine des données

Recommandation :
Précisez le mode de collecte et l'origine des données
Modèle / exemple :
Collecte de données primaires avec la numérisation de sources papier - précisez les centres d'archives, bibliothèques, centres d'études supérieures, …
Moissonnage de ressources en ligne - précisez l'origine des données (Gallica, Europeana, autre bibliothèque numérique)
Collecte et extraction de passages et extraits de sources (textes, images) - précisez les sources...

Période couverte par le corpus, auteur(s) concerné(s) et organisation du corpus

Recommandations :
Décrivez l'organisation du corpus, l'arborescence des fichiers, le système de nommage et de gestion des répertoires et des fichiers.
Vous pouvez vous référer aux recommandations en ligne sur https://doranum.fr/stockage-archivage/comment-nommer-fichiers/. Y sont spécifiées les 5 règles de nommage essentielles des fichiers et documents.

Etat du corpus numérique

Recommandations :
Indiquez la nature des données, leurs formats, leur volumétrie (en poids et nombre de fichiers), leur état (le corpus corpus est-il complet ? Si non pour quelles raisons ? L'état physique des sources était-il altéré, ce qui peut impacter sur la qualité de leur version numérique ?)
Pour que les données soient réutilisables sur le long terme, assurez-vous d'utiliser des formats ouverts, non propriétaires et d'un usage répandu au sein de la communauté de recherche. Pour vérifier l'éligibilité de vos formats vous pouvez utiliser l'outil https://facile.cines.fr/
Modèle / exemple :
Nom du jeu de données : Environ 16 000 feuillets en bon état de conservation, numérisés et"océrisés", qui représentent un poids d'environ 1,7Go (sur un corpus de 100 000 pages au total qui sera numérisé pour un total estimé de 16Go). Les données sont disponibles en mode texte au format PDF.
Nom du jeu de données : 950 images au format jpeg (env. 7Go)

Modifications effectuées sur les données, versions

Recommandations :
Les données ont-elles subi des traitements ? Si oui lesquels, par quels moyens et avec quels outils ? Retracez l'historique de ces modifications.
Soyez"fair": pour garantir l'accessibilité et la réutilisation des données il est recommandé de privilégier des logiciels libres de droits et open source.

Autres données créées ou collectées pour documenter et/ou enrichir les corpus constitués.

Recommandations :
Précisez le type et la nature de ces données. Il peut s'agir de notices descriptives ou bibliographiques, données d'analyse quantitative effectuées sur des textes, données issues de modélisation ou de simulation, …
Indiquez le mode de création ou de collecte (en précisant leur origine pour les données réutilisées) de ces données. L'objectif de leur création ou de leur collecte. Leur volumétrie.

Métadonnées préexistantes, métadonnées créées et standards et formats utilisés

Recommandations :
Il est recommandé d'utiliser un schéma de description courant, à l'instar du Dublin Core qui est également un standard interdisciplinaire. Il est également possible d'utiliser des standards spécifiques au type de données et à la discipline (EAD pour la description d'archives, MARC ou UniMARC pour des données bibliographiques, …).
Les formats d'échanges les plus courants sont l'XML et le CSV.
Pour s'informer sur les métadonnées, standards et formats : https://doranum.fr/metadonnees-standards-formats/fiche-synthetique/

Référentiels d'indexation et vocabulaires contrôlés, thésaurus ou ontologies disciplinaires utilisés

Documentation destinée à accompagner les métadonnées en vue de la réutilisation des données

Recommandations :
L'ANR spécifie dans son modèle de PGD que "la documentation accompagnant les données permet aux utilisateurs de les repérer facilement et apporte les informations nécessaires à un bon usage et une bonne interprétation". Il peut s'agir a minima d'un fichier "read me" rassemblant les informations générales sur les données, de la documentation destinée à la description et à la compréhension de l'organisation des données, un glossaire pour le vocabulaire spécifique et les acronymes, etc.

4) Stockage, sauvegarde et sécurité des données

Documentation numérique ou papier décrivant et renseignant le lieu de stockage final, les lieux et infrastructures de stockage des données pendant le projet

Recommandations *:*
Décrivez ici le matériel physique et les lieux de stockage des données.
Stockez vos données dans au moins 2 endroits. Évitez le stockage externe et privilégiez les outils mis à disposition par l'institution.
Les bonnes questions à se poser pour organiser au mieux la sauvegarde des données[7] *:*

- *Quel volume approximatif de données devrons-nous sauvegarder ? L'espace de stockage est-il suffisant ?*
- *Quelle sera la périodicité des sauvegardes ? quotidienne ? hebdomadaire ? mensuelle ?*
- *Quel service utiliser, fourni par quelle institution ?*
- *Les données sont-elles sous contrat de maintenance ?*
- *Faudra-t-il accéder fréquemment aux données ? en temps réel ?*
- *etc.*

Volumétrie des données stockées. Modalités de sauvegarde et de protection des données

Recommandations :
Indiquez ici les procédures de sauvegarde mises en place (fréquence des sauvegardes ? Automatisée ou non ?), les personnes en charge de la protection des données et du contrôle de l'accès, le mode de récupération des données en cas d'incident…
Indiquez également si l'accès aux données est sécurisé ou non.

Risques

Recommandations :
Différents facteurs sont susceptibles de menacer l'intégrité, la disponibilité et la confidentialité des données produites au cours du projet. Les risques peuvent être de différentes natures : des risques naturels pesant sur les infrastructures (zones sismiques, inondables etc.), des risques techniques (corruptions ou pertes de données, problèmes de serveurs etc.), des risques de confidentialité (accès non autorisés, fuites de données sensibles, etc.)[8]

[7] D'après : Atelier Données,"Guide de bonnes pratiques sur la gestion des données de la recherche", janv. 2021. https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html (consulté le juill. 13, 2021).
[8] D'après A.CARTIER, R.DELEMONTEZ, M.MOYSAN, N.REYMONET. Réaliser un plan de gestion de données"FAIR": guide de rédaction (v2, 2018)

5) Accès et partage des données

Modalités d'accès et de partage des données pendant la durée du projet

Recommandations :
Précisez si des données seront d'emblée entièrement accessibles sans restriction.
Existe-t-il des contraintes ou des modalités de diffusion particulières pour certains jeux de données (période limitée de partage, embargo) ?
Indiquer si les données sont accessibles selon des restrictions ou des conditions particulières : sont-elles sur un serveur local, un intranet, sur internet, en libre accès ou avec un accès authentifié. Préciser, le cas échéant, les différents niveaux d'habilitations et les règles qui les régissent.
Modèle / exemple :
Les données primaires seront rendues publiques en accès libre, à partir de textes et d'images intégralement libres de droits.
Les métadonnées seront disponibles sous la forme d'apparat critique et de notes explicatives (éditions scientifiques intégrales, dossiers d'études, données bibliographiques, ressources informatives historiques). Elles seront rendues publiques sous la réserve du respect de la propriété intellectuelle de leurs auteurs (professeurs, chercheurs, spécialistes universitaires).

Limites éventuelles à l'accès aux données

Recommandations :
Utilisation nécessaire de logiciels propriétaires par exemple.

Partage des données

Recommandations :
Indiquez par exemple si elles sont stockées sur un entrepôt de données, indexées dans un catalogue, accessibles par demande directe…
Dans le cadre du consortium CAHIER et afin de respecter les principes FAIR, le consortium recommande le stockage des données sur Nakala, le service de dépôt et de stockage sécurisé proposé par la TGIR HumaNum pour les SHS. Un second dépôt sur un autre entrepôt de votre choix ou sur une plateforme institutionnelle reste possible dans tous les cas.
Modèle / exemple :
Toutes les données sont libres de droits. Elles seront accessibles sur une interface dédiée et déposées sur Nakala. Les données et métadonnées seront interopérables et moissonnables par l'intermédiaire du protocole OAI-PMH.

6) Responsabilités et ressources pour la gestion des données

Identifiez et décrivez les rôles de responsabilité des données dans votre projet, et nommez si possible les personnes impliquées.

Recommandations :
Nommez la/les personnes responsables de la saisie des données, de la production des métadonnées, du traitement, de l'analyse, du stockage et de la sauvegarde des données ainsi que de leur partage et éventuellement leur archivage. Un responsable de la coordination du système de gestion des données pourra être nommé. Celui-ci devra idéalement être impliqué dans le pilotage du projet.
Modèle / exemple *:*
Responsable de la gestion des données : Nom prénom, institution, ville, adresse mail.
Responsable de la qualité des données (traitement, anonymisation, format, nettoyage, …) : Nom prénom, institution, ville, adresse mail
Responsable de la collecte des données : Nom prénom, institution, ville, adresse mail
Responsable du stockage et de la sécurité des données : Nom prénom, institution, ville, adresse mail

Évaluez les coûts (budgets, personnels et temps) dédiés à rendre les données FAIR (temps et budgets pour la collecte et la diffusion des données, pour le stockage et l'archivage).

Recommandations :
Moyens matériels (serveurs, machines virtuelles, logiciels dédiés, licences supplémentaires, …)
Moyens humains (recrutement évalué en ETP pour des tâches spécifiques relevant de la gestion des données, actions de formation, …)

7) Archivage des données

Quelles sont les données à conserver sur le moyen et le long terme ? Quelles sont les données à détruire ?

Recommandations :
A l'issue du projet, des jeux de données se prêteront à une conservation à long terme pour une utilisation future, tandis que d'autres données ne nécessitent qu'une préservation à moyen terme car jugées moins essentielles et au potentiel de réutilisation limité, voire destructibles pour des raisons légales et de confidentialité. Expliquez le choix des données à conserver et à détruire.

Sur quelle plateforme est prévu l'archivage pérenne des données ? Si un autre moyen est envisagé, précisez lequel et décrivez les outils et méthodes.

Recommandations :
Indiquez quelle plateforme pourrait accueillir les données et à quel coût.
Certaines institutions proposent un service d'archivage des données de la recherche. Actuellement le service de référence mandaté par le Ministère de l'enseignement supérieur et de la recherche est le CINES.

Durée de conservation des données

Recommandations :
Elle est à déterminer en fonction des jeux de données, des coûts alloués à l'archivage et au service de conservation choisi.

Volume des données à conserver

Modèle / exemple :
L'ensemble des données est voué à être conservé sur le long terme. Ce qui représente environ 50 Go.

Coûts alloués à la conservation

Recommandations :
Les coûts nécessaires de l'archivage sont à déterminer en fonction de la durée de conservation de vos données, du volume à conserver et de la plateforme sélectionnée.

Quels outils, méthodes, procédures seront nécessaires pour accéder à ces données archivées et les réutiliser ? (logiciels spécifiques, identification et droits pour accéder à la plateforme, ...)

8) Partage des données à l'issu du projet

Politique de dissémination des données

Recommandations :
Précisez s'il existe des limites à la diffusion des données. Expliquez également comment les données pourront être trouvées et réutilisées par les pairs, voire par le grand public. En effet, les données devront dans la mesure du possible respecter des principes FAIR, c'est-à-dire être trouvables (findables), accessibles, interopérables et réutilisables : les données devront être faciles d'accès, identifiables et citables grâce à des identifiants uniques (DOI), leur usage est facilité a minima par l'accompagnement d'une description et de documentation, des formats ouverts et non propriétaires, un lieu de stockage (entrepôt) ouvert, gratuit et référencé par les moteurs de recherche.

Potentiel de réutilisation des données

Recommandations :
Expliquez ici le potentiel de réutilisation des données et à qui elles pourraient être utiles. Il est notamment possible de suggérer d'autres axes de recherche à partir du projet
Modèle / exemple :
Les données publiées visent à documenter la recherche internationale sur la circulation de la pensée durant la première moitié du vingtième siècle. En raison du contexte de ces exils (guerres d'Espagne, guerres mondiales), il reste encore des zones d'ombres sur l'itinéraire et le parcours de certains intellectuels de l'époque. La publication de ces données vise à apporter des informations permettant d'éclaircir ces zones d'ombre.
Ces données s'adressent avant tout à un public de chercheurs en sciences humaines et sociales, mais peuvent également intéresser un public d'amateurs.

Eléments d'accompagnement qui permettent la réutilisation des données.

Recommandations :
Indiquez si de la documentation accompagne les jeux de données afin d'en faciliter la compréhension et réutilisation à long terme : des documents permettant de les décrire, d'en expliquer l'usage et les potentialités d'usage, la manière dont les données ont été produites ou collectées, quelles sont les contraintes d'usage, … voire de les enrichir.
Modèle / exemple :
Thésaurus, ontologie, codes informatiques, algorithmes, document explicatif (read me), inventaire, bibliographie,...

Publications sur les données pour en améliorer l'exposition

Recommandations :
Publications faites sur le projet et l'exploitation de ces jeux de données.
La rédaction d'un data paper est généralement recommandée. On peut également citer des articles publiés sur le le projet et faisant des liens avec les données.

Conditions de réutilisation (licences et contrats pour l'ensemble du projet et sur chaque jeu de données)

Recommandations :

Donnez les conditions de réutilisation des données pour le projet.

Listez pour chaque jeu de données : les conditions et modes d'accès aux données (accès libre ou restreint, sur autorisation, selon le statut, …), s'il est sous licence ou contrat[9], s'il est contraint par un embargo ou d'accès restreint pour des questions de confidentialité ou de sensibilité des données.

Modèle */* exemple :

Les données initiales sont dans le domaine public. Certaines données créées dans le cadre du projet sont librement accessibles par téléchargement sous une licence Attribution – Pas d'Utilisation Commerciale – Partage dans les Mêmes Conditions 4.0".

[9] Voir **A.CARTIER, R.DELEMONTEZ, M.MOYSAN, N.REYMONET**. *Réaliser un plan de gestion de données"FAIR": guide de rédaction* (v2, 2018), p31.
Disponible en ligne : https://archivesic.ccsd.cnrs.fr/sic_01690547v2/document
"La licence précise les conditions de partage et de réutilisation des données diffusées dans le cadre du projet, ainsi les éventuelles contreparties intellectuelles ou économiques qui y sont associées. Il est important de préciser qu'une diffusion en libre accès ne signifie pas nécessairement qu'une œuvre est libre de droit. La licence a notamment pour objectif de clarifier le statut juridique d'une œuvre et de préciser les conditions d'usage. De trop nombreux contenus ne sont pas réutilisés au maximum de leur potentialité en raison des ambiguïtés juridiques dues à l'absence de licences explicites. Il existe de nombreuses licences libres à utiliser selon les législations, les formats de données, les souhaits de protection des auteurs, les exigences des financeurs, etc."
Pour les jeux de données (liste non exhaustive)

- CC-by 4.0 : https://creativecommons.org/licenses/by-sa/4.0/deed.fr
- CC0 : https://creativecommons.org/publicdomain/zero/1.0/legalcode.fr
- La licence ETALAB : https ://www.etalab.gouv.fr/licence-ouverte-open-licence

Pour les bases de données (liste non exhaustive)

- OKF – Open Database License (OdbL) : http://opendatacommons.org/licenses/odbl/1.0/
- OKF – ODC-By : http://opendatacommons.org/licenses/by/1.0/
- OKF – ODC-PDDL (Public domain) : http://opendatacommons.org/licenses/pddl/1.0/

Pour les logiciels (liste non exhaustive)

- Open Licence Software (OPS) : https://opensource.org/licenses/OSL-3.0
- GNU-GPL : http://www.gnu.org/licenses/gpl.html