Guide pour la FAIRisation des données des corpus d'auteurs préparé par Fatiha Idmhand et Ioana Galleron V2 [Groupe de travail Data_Cahier]

Fatiha Idmhand, Ioana Galleron

[Groupe de travail Data_Cahier]

Guide pour la FAIRisation des données des corpus d'auteurs

06-09-2020

Les principes FAIR (Findability, Accessibility, Interoperability and Reusability) définissent un ensemble minimal de principes qui permettent aux machines et aux humains de trouver, d'accéder, d'interopérer et de réutiliser les données et métadonnées de recherche. **Les principes FAIR doivent être considérés comme des bonnes pratiques destinées à faciliter la réutilisation des données et des résultats de la recherche.**

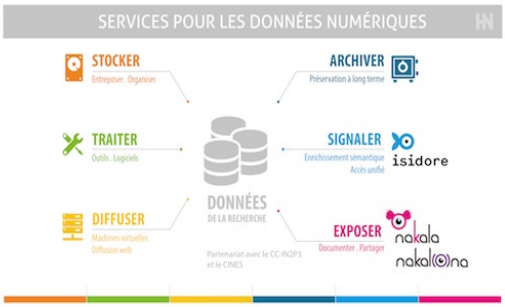
Le consortium CAHIER, et ses membres, recommandent et mettent en œuvre ces bonnes pratiques.

Dans le cycle de vie d'un projet, le dépôt des données (4) en vue de leur archivage arrive généralement en fin de chaîne, après la collecte (1), le traitement (2) et l'analyse (3). Le dépôt contribue à rendre pérennes et à mettre en valeur les résultats. Pour chaque phase du cycle de vie d'un projet, l'infrastructure Huma-Num propose différents outils :

NAKALA permet à des équipes de recherche, qui en font la demande, de déposer leurs données numériques (fichiers texte, son, image, vidéo) dans un entrepôt sécurisé qui assure à la fois l'accessibilité aux données et leur citabilité dans le temps.

Les technologies mises en oeuvre permettent notamment de rendre interopérables les métadonnées, c'est-à-dire la possibilité de pouvoir les connecter à d'autres entrepôts existants, et de les rendre moissonnables par des services spécialisés comme *ISIDORE*.

NAKALA s'inscrit dans un dispositif cohérent de services mis en place par la TGIR Huma-Num pour faciliter l'accès, le signalement, la conservation et l'archivage à long terme des données numériques de la recherche en SHS.



Le consortium CAHIER utilise Nakala pour stocker ses données afin qu'elles soient trouvables, accessibles, interopérables et réutilisables. Ce petit guide décrit le processus à mettre en œuvre pour rendre les données du Consortium CAHIER « FAIR ». Celui-ci est organisé en quatre étapes :

1° Évaluer le degré d'ouverture de ses projets
2° Confronter ses métadonnées aux attentes du consortium
3° Compléter et corriger les métadonnées si nécessaire
4° Déposer sur Nakala, obtenir un identifiant pérenne et l'associer aux documents publiés sur son propre site web (ou sur un site web institutionnel).

1° Mon projet est-il FAIR ?

Pour être FAIR, les données produites dans le cadre du projet doivent être trouvables (Findable), accessibles et téléchargeables librement (Accessible), interopérables (Interoperable) et réutilisables (Reusable). Voici quelques éléments concrets liés à cet objectif, et auxquels le dépôt sur Nakala apporte une réponse.

F

Pour que les données et métadonnées soient trouvables, il faut les pourvoir d'un identifiant pérenne et unique au niveau mondial de type DOI, Handle ou Ark par exemple. Il convient également d'être certain que ses métadonnées puissent être moissonnées par les agrégateurs des grandes bibliothèques numériques.

Si votre institution de rattachement ou un partenaire clé de votre projet (bibliothèque, service d'archives ou informatique, etc.) offre de tels services et qu'elle donne à vos données et métadonnées un identifiant de type DOI, Handle ou Ark, alors les données sont trouvables (Findable) et la première condition est remplie. Toutefois, Nakala reste utile d'effectuer le dépôt sur Nakala et d'équiper vos données de deux identifiants car Nakala offre une série de services complémentaires.

Dans la majorité des cas, les institutions hébergeant des projets (ou les sites webs de projets) ne proposent pas de DOI, ARK ou Handle. Vos données sont visibles via le site web hébergé sur le serveur d'université mais sans identifiants uniques et pérennes, et sans garantie de visibilité. Le dépôt sur Nakala est, dans ce cas, une nécessité.

A

Construire, fournir, livrer un site web ne répond que partiellement à la question de l'Accessibilité et cela, même si le site est « hébergé par » ou « chez » Huma-Num car vos données ne sont pas, pour autant, accessibles au même niveau que des données versées dans un entrepôt ouvert comme Nakala. Force est également de constater que dans de nombreux cas, une grande partie des ressources n'est pas accessible sur le site web : conditionner l'accès aux données par une demande d'inscription préalable ou imposer la consultation par l'intermédiaire exclusif d'une interface va à l'encontre de l'objectif de l'Accessibilité.

Si la non-exposition des données peut être nécessaire dans la phase de préparation des documents, il est essentiel qu'au terme de l'existence du consortium CAHIER une partie significative de la volumétrie envisagée au moment de la labellisation du projet soit accessible. Pour les images protégées (p. ex., facsimilés de manuscrits d'un auteur contemporain, images obtenues en accord de partenariat avec une bibliothèque, etc.), cette accessibilité peut ne concerner que les métadonnées des documents et, idéalement, les transcriptions.

Les membres de CAHIER sont d'ailleurs invités à engager un dialogue avec les auteurs, leurs ayant-droit, ou avec les institutions qui leur fournissent les images : une explication de l'intérêt de l'open access permet parfois d'obtenir leur accord pour une plus ample exposition des données. Les institutions patrimoniales françaises ont d'ailleurs leurs propres obligations en termes d'accès ouvert, mais ne connaissent pas forcément l'existence de Nakala.

I

La préparation des données selon des pratiques et des référentiels mondialement connus et partagés est essentielle pour assurer leur interopérabilité. Dans votre propre projet, vous avez probablement déjà veillé à cet aspect, par exemple en utilisant un CMS[1] pour saisir et exposer vos données. Le dépôt sur Nakala vient dans le prolongement de cet effort et renforce cet aspect.

R

La réutilisabilité a concerné, jusqu'à présent, la qualité du fichier numérique et son format : ouvert, standardisé, etc. Néanmoins, les conditions de cette réutilisabilité ont été moins étudiées et pensées. Le dépôt d'un grand volume de données sur une même plateforme est à même de stimuler cette réutilisabilité, en donnant plus de visibilité à votre projet.

F+A+I+R

Le consortium CAHIER utilise nakala pour remplir l'ensemble des conditions FAIR pour ses données.

2° Confronter ses métadonnées aux attentes du consortium

Le RDA FAIR Data Maturity Model Working Group dans sa version d'avril 2020 a classé, en fonction de leur statut plus ou moins prioritaire, les indicateurs des données FAIR. Après avoir consulté ces indicateurs dans leur version d'avril 2020[2], et en vue d'harmoniser les pratiques de dépôt sur nakala, le consortium a défini un modèle minimal commun des métadonnées qui doivent équiper les fichiers produits dans le cadre du consortium.[3]

Le tableau ci-dessous décrit les correspondances entre deux standards habituellement utilisés dans les projets (TEI et Dublin Core) d'une part, et le standard utilisé par Nakala[4] (DCTerms), d'autre part.[5]

L'objectif ici n'est pas de définir un teiHeader commun pour les différents projets des membres du consortium CAHIER ou une liste finie de champs Dublin Core. **Ne figurent dans ce qui suit que les métadonnées indispensables** exprimées d'une part sous forme de balises dans un header TEI, et d'autre part sous forme de métadonnées Dublin Core. En d'autres mots : **il peut y avoir d'autres éléments dans le header/les notices, mais il faut a minima faire figurer ceux-ci.**

Ce socle commun de métadonnées descriptives peut être étendu ad libitum mais il est utile pour la création de filtres d'export et afin de faire converger les pratiques d'encodage.[6]

[1] Omeka, Wordpress, etc.
[2] RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. Research Data Alliance. DOI: 10.15497/RDA00045. Voir : https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines
[3] Les recommandations décrites ci-après peuvent être réutilisées et appliquées par d'autres projets sur des Corpus d'auteurs.
[4] https://documentation.huma-num.fr/content/14/86/fr/vademecum-nakala.html
https://isidore.science/document/10670/1.8guiu4
[5] Pour plus d'informations concernant les attentes de Nakala, consultez https://humanum.hypotheses.org/5989.
[6] Il est également recommandé de se reporter au fichier xml teiHeader-modele-data-cahier, qui donne des explications plus détaillées pour chacun des champs.

Guide métadonnées pour Nakala

Document préparé par Nathalie Arlin (2018), relu, révisé et complété par Ioana Galleron et Fatiha Idmhand (12-05-2020)

Champ requis	DC/DCterms	TEI	Exemple
Titre	dc:title / dcterms:title	teiHeader/fileDesc/titleStmt/title @type="main"	Exemple 1 : Le Rouge et le violet Exemple 2 : Le Rouge et le violet : édition électronique Exemple 3 : Lettre de Pierre à Paul Exemple 4 : [Ceci est un titre forgé. Le document n'a pas de titre, j'en donne un et je le mets entre crochets. L'usage veut que l'on reprenne, normalement, la première phrase du manuscrit ou de la lettre dans le cas de correspondances.]
Auteur	dc:creator / dcterms:creator	teiHeader/fileDesc/titleStmt/author	Exemple 1 : Sartre, Jean-Paul Exemple 2 : Zola, Emile NB. Le champ auteur peut être dupliqué.
Éditeur scientifique	dc:contributor / dcterms:contributor	teiHeader/fileDesc/titleStmt/editor	Exemple 1 : Dupont, Jeanne (Professeur des Universités) Exemple 2 : Dupont, Jeanne (Professeur des Universités) | Itterom, Ocnar (Chercheur CNRS) Exemple 3 : Ghog, Nav (Critique d'art)
Éditeur	dc:publisher / dcterms:publisher	teiHeader/fileDesc/publicationStmt/publisher	Exemple 1 : Projet Région n°12345 | Consortium CAHIER TGIR Huma-Num Exemple 2 : Projet FLG – AAP MSH Centre Sud NB : Le champ publisher peut être dupliqué.
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 Ou teiHeader/fileDesc/publicationStmt/availability/licence/date (si présent)	Exemple 1 : 2020-07-14 NB : Date de « délivrance » officielle de la ressource numérique. Ne pas confondre avec la date de création (voir ci-après)
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" Ou teiHeader/profileDesc/creation/date @notBefore="1700" @notAfter="1750"	Il s'agit de la date de création du document source (appelée date première dans le catalogue OAI du consortium CAHIER) Exemple 1 : 1923-11-02 Exemple 2 : 1854-03 Exemple 3 : 1765 Exemple 4 : [1920-1940]

Informations sur les droits d'utilisation	dc:rights / dcterms:rights	teiHeader/fileDesc/sourceDesc/bibl/note[@type='settlement'] ou teiHeader/filedesc/sourceDesc/biblStruct/note[@type='settlement'] ou, pour les manuscrits, teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/ settlement	Exemple 1 : Fonds Jean-Charles Carpo – Bibliothèque Jacques Tecuod Exemple 2 : Archives familiales Jean-Sol Ertrap
Identifiant	dc:identifier / dcterms:identifier NakalaNakala	teiHeader/fileDesc/publicationStmt/idno[@type="URL"] Dans un second temps on ajoutera teiHeader/fileDesc/publicationStmt/idno @type="DOI"	NB. Le protocole OAI requiert que chaque ressource soit identifiée par une URI qui renvoie au fichier xml et non à l'adresse du site. Il est recommandé d'utiliser un idno de type cote de bibliothèque avant le dépôt puis, dans un second temps, l'ajout d'un idno de (@type="DOI"). L'identifiant pérenne sera délivré après le dépôt sur Nakala. Les fichiers seront (re)mis à jour pour être renseignés grâce au DOI.
URI de licence	dc:rights / dcterms:license Nakala	teiHeader/fileDesc/publicationStmt/availability/licence[@target="URI de la licence"]	Exemple 1 : CC-BY 4.0 Exemple 2 : CC-BY-NC-SA 4.0 Recommandation : Publier l'information de licence sous forme d'URI, d'autant plus s'il s'agit d'un type de licence connu (par exemple Creative Commons) Nakala recommande les licences Creative Commons à choisir parmi : CC-BY 4.0 ; CC-BY-SA 4.0 ; CC-BY-ND 4.0 ; CC-BY-NC 4.0 ; CC-BY-NC-SA 4.0; CC-BY-NC-ND 4.0
Référence (bibliographique) du document d'origine	dc:source / dcterms:source	teiHeader/fileDesc/sourceDesc/bibl/note[@type='identifier'] ou pour une bibliographie plus détaillée : teiHeader/fileDesc/sourceDesc/biblStruct/note[@type='identifier'] Pour les manuscrits teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/[différents champs pertinents, don't idno]	On mentionnera ici les données bibliographiques du document source (date, lieu et l'année de publication de la source, cote du document dans l'institution) Exemple 1 : NAF 10266 Exemple 2 : Cote : NAF 10266

Langue	dc:language / dcterms:language	teiHeader/profileDesc/langUsage/language @ident="fr"	Exemple 1 : fr Exemple 2 : es Répétable pour chaque langue – norme iso 639-2b. Il est recommandé de privilégier le code à 2 caractères quand il est suffisamment précis. Cf: https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
Résumé	dc:description / dcterms:description	teiHeader/profileDesc/abstract	Exemple 1 : Le roman parle de …. Exemple 2 : Ce document est le premier de… Présentation du texte à destination du public, si ce champ n'est pas renseigné, seront extraites les premières lignes du plein texte du document
Mots-clés	dc:subject / dcterms:subject	teiHeader/profileDesc/textClass/keywords/term @type="subject"	Liste de mots-clés destinés à l'indexation thématique ou générique Exemple 1 : Exil | Guerre Exemple 2 : Renaissance | Philosophie Exemple 3 : Roman | Prose | Première personne

Type de fichier dépos	dc:format / dcterms:type	Par défaut, tous les fichiers TEI seront déclarés comme “Text”	Il s’agit du type de présentation matérielle ou numérique du contenu de la ressource. La pratique recommandée est de choisir une valeur dans un vocabulaire contrôlé (par exemple, dans la liste du DCMI Type Vocabulary). Ci-dessous une liste des correspondances entre les extensions de fichier et les variantes acceptées des “intitulés” du “Type” à renseigner en fonction du fichier de la ressource. Si extension du fichier : 'doc', 'xls', 'xlsx', 'pdf', 'docm', 'docx', 'epub', 'gdoc', 'htm', 'html', 'log', 'mcw', 'odm', 'odoc', 'odt', 'ott', rtf', 'txt' alors type = text, doc, pdf Si extension du fichier : 'png', 'jpg', 'jpeg', 'tif', 'tiff', 'gif', 'bmp', 'webp', 'svg', 'djvu', 'ico', 'icns', 'jng', 'iff', 'jp2', 'jps', 'mng', 'pbm', 'pnm', 'ppm', 'tga', 'tiff', 'xpm' alors type = image, stillimage Si extension du fichier : 'mpg', 'avi', 'mkv', 'mp4', 'webm', 'flv', 'vob', 'ogv', 'avi', 'mov', 'qt', 'mts', 'm2ts', 'wmv', 'yuv', 'rm', 'rmvb', 'asf', 'amv', 'm4p', 'm4v', 'mp2', 'mpeg', 'm2v', 'm4v', '3gp', '3g2' alors type = vidéo, video, movingimage Si extension du fichier : 'mp3', 'wav', 'flac', 'gsm', 'm4a', 'ogg', 'oga', 'mogg', 'ra', 'vox', 'wma' alors type = sound,audio

Informations recommandées			
Documents en relation	dc:relation / dcterms:relation	teiHeader/sourceDesc/bibl/note[@type='relation'] **ou** teiHeader/sourceDesc/biblStruct/note[@type='relation']	Exemple 1 :<note>Cette édition est la version pirate de la précédente</note> Exemple 2 : Le rose et le jaune – Manuscrit 1 | Le rose et le jaune – Manuscrit 3 | Le rose et le jaune – Carnet 1 Exemple 3 : NAF 10266 | NAF 10267 | NAF 10268
Autres versions (autres versions faisant foi)	dc:isVersionOf / dcterms : dc:isVersionOf	teiHeader/fileDesc/editionStmt/edition (pour une édition no. 2 prenant appui sur une E1) ou teiHeader/encodingDesc/projectDesc (pour les adaptations, les versions différentes, etc.)	La ressource décrite est une version, une nouvelle édition, ou une adaptation de la ressource référencée. Des changements de versions impliquent de réels changements du contenu plutôt que des différences dans le format. Exemple 1 : <edition> Première édition <date>2017</date> </edition> Exemple 2 : Le rose et le jaune – Manuscrit 2
Informations propres à Nakala et recommandées			
Description spécifique Nakala **Droits**	Deux options : accès libre / accès protégé		Exemple 1 : accès libre Exemple 2 : accès protégé | Accès restreint jusqu'au 2023-21-31 en raison de….
Description Spécifique Nakala **Collection**	C'est là que l'on peut reconstruire les regroupements intellectuels. Suggestion : remplir par défaut ce champ avec l'information contenue dans le champ Titre		Exemple 1 : Le rouge et le violet Exemple 2 : [Dossier préparatoire xxxxx] Exemple 3 : [Projet Partre numérique]