FAIRiser vos données : mémorandum

Fatiha Idmhand, Ioana Galleron, Ala Eddine Laouir, Andres Echavarria, Laurent Passion

FAIRiser vos données : mémorandum

consortium cahier

Andrés ECHAVARRÍA
Ala Eddine LAOUIR
Fatiha IDMHAND
Ioana GALLERON
Laurent PASSION

Guide pour la FAIRisation des données des corpus d'auteurs

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 de bonnes pratiques destinées à faciliter la réutilisation des données et les 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 HumaNum propose différents outils.

Le consortium CAHIER utilise principalement 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.

Services Huma-Num
par étapes

on des données

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 vos données sont trouvables (Findable) et la première condition est remplie. Toutefois, il reste utile d'effectuer le dépôt sur Nakala et d'équiper vos données de deux identifiants car l'outil offre une série de services complémentaires.

Dans la majorité des cas, les institutions hébergeant des projets (ou les sites web de projets) ne proposent pas de DOI, Handle ou ARK. Ainsi, même si vos données sont visibles via le site web hébergé sur le serveur de votre institution, vos données ne sont pas FAIR en l'absence d'identifiants uniques et pérennes. Nous vous recommandons de doubler cette publication numérique d'un dépôt sur un entrepôt de données comme Nakala. C'est même une nécessité de la science ouverte.

A

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. 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 d'un projet, une partie significative des données soit accessible. Lorsque des images ne peuvent être rendues, un embargo est possible sur ces données, toutefois, cette restriction peut ne pas concerner les métadonnées ou les transcriptions. Des données peuvent donc être mises en ligne.

Les membres de CAHIER sont invités à engager un dialogue avec les auteurs,

leurs ayants-droits, 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.

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

Guide pour la FAIRisatio

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 pour saisir et exposer vos données. Le dépôt sur Nakala vient dans le prolongement de cet effort : après l'exposition, il vous reste à stocker vos données.

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.

2° Confronter ses métadonnées aux attentes du consortium

Le RDA FAIR Data Maturity Model Working Group a publié en avril 2020 un guide comportant des indicateurs des données FAIR. Après avoir consulté ces indicateurs, et en vue d'harmoniser les pratiques de dépôt, le consortium a défini un modèle minimal commun des métadonnées qui doivent accompagner les fichiers produits dans le cadre du consortium.

Ce socle commun de métadonnées descriptives peut être étendu ad libitum mais il est utile de faire converger les pratiques d'encodage.

Il est important d'aborder la question de la structuration des métadonnées : l'identification du format utilisé pour les structurer permettra d'organiser le dépôt en masse des données dans l'entrepôt.

Le consortium CAHIER a développé une application qui communique avec la plateforme Nakala à travers son API.

FAIR Data Maturity Model: specification and guidelines

Plus d'information sur Nakala

Au mois de décembre 2020, la nouvelle version de la plateforme de stockage de données Nakala a été présentée par Huma-Num. Cette nouvelle version a intégré une interface utilisateur plus facile à utiliser, ce qui a été très apprécié par les utilisateurs. De plus, l'API de Nakala est devenue plus riche, facilitant ainsi l'intégration de toutes les fonctionnalités de Nakala dans d'autres outils.

Nakala utilise un triplestore pour enregistrer les données fournies par les utilisateurs. Cette méthode de sauvegarde facilite la publication des données sur le Web et dans le monde de l'Open Data (ou Linked Data). Ces services seront très intéressants pour les chercheurs à l'avenir.

Nakala n'a pas encore développé le service de dépôt en masse des données. Comme ce manque pouvait limiter les projets de CAHIER, le consortium a créé une application web, *mynkl*, qui communique avec l'API de Nakala et qui facilite le dépôt de grands volumes de données. L'appilcation *mynkl* comble, provisoirement, un manque.

Plus d'information sur ISIDORE

mynkl

L'outil mynkl

Services fournis par nakala

Interface Nakala

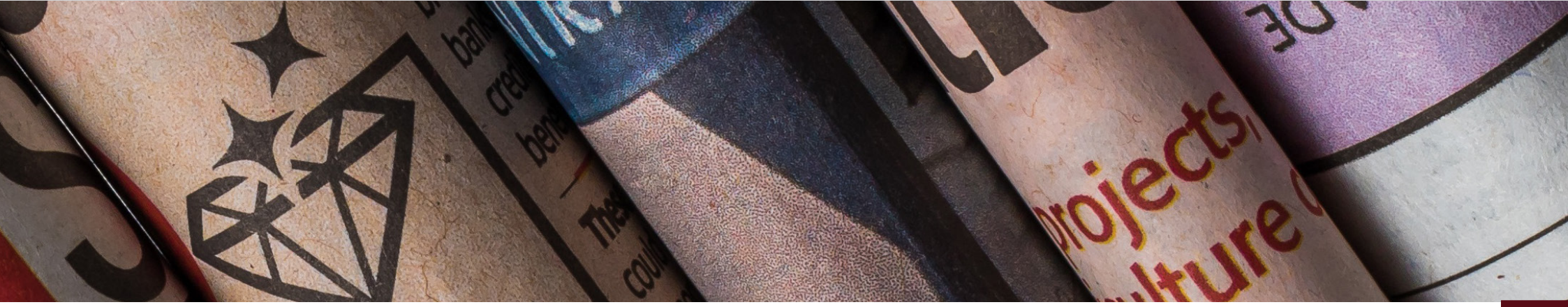
Base de données RDF ou Triplestore

Point d'accès SPARQL Nakala

API Nakala

Chercheur

Outil MyNkl développé par consortium cahier

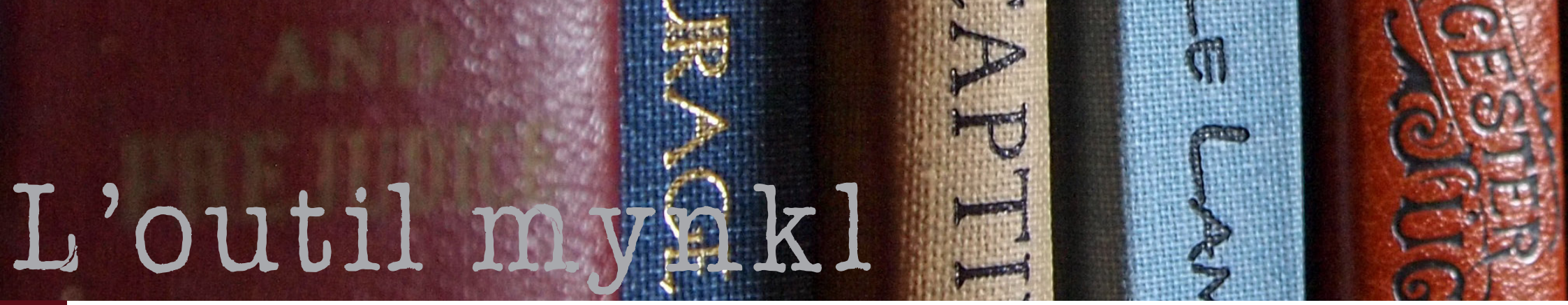


Quelles métadonnées utilise l'outil mynkl ?

Les projets du Consortium CAHIER ont principalement utilisé le Dubln Core, le XML-TEI ou des tableurs (csv) pour structurer leurs métadonnées.
L'entrepôt de stockage de données Nakala utilise le standard DCTerms pour structurer les métadonnées des objets déposés.
Pour faciliter le dépôt des données, le Consortium a puisé dans les teiHeader et les balises Dublin Core des projets (ou dans les tableurs), les informations nécessaires au dépôt dans Nakala.

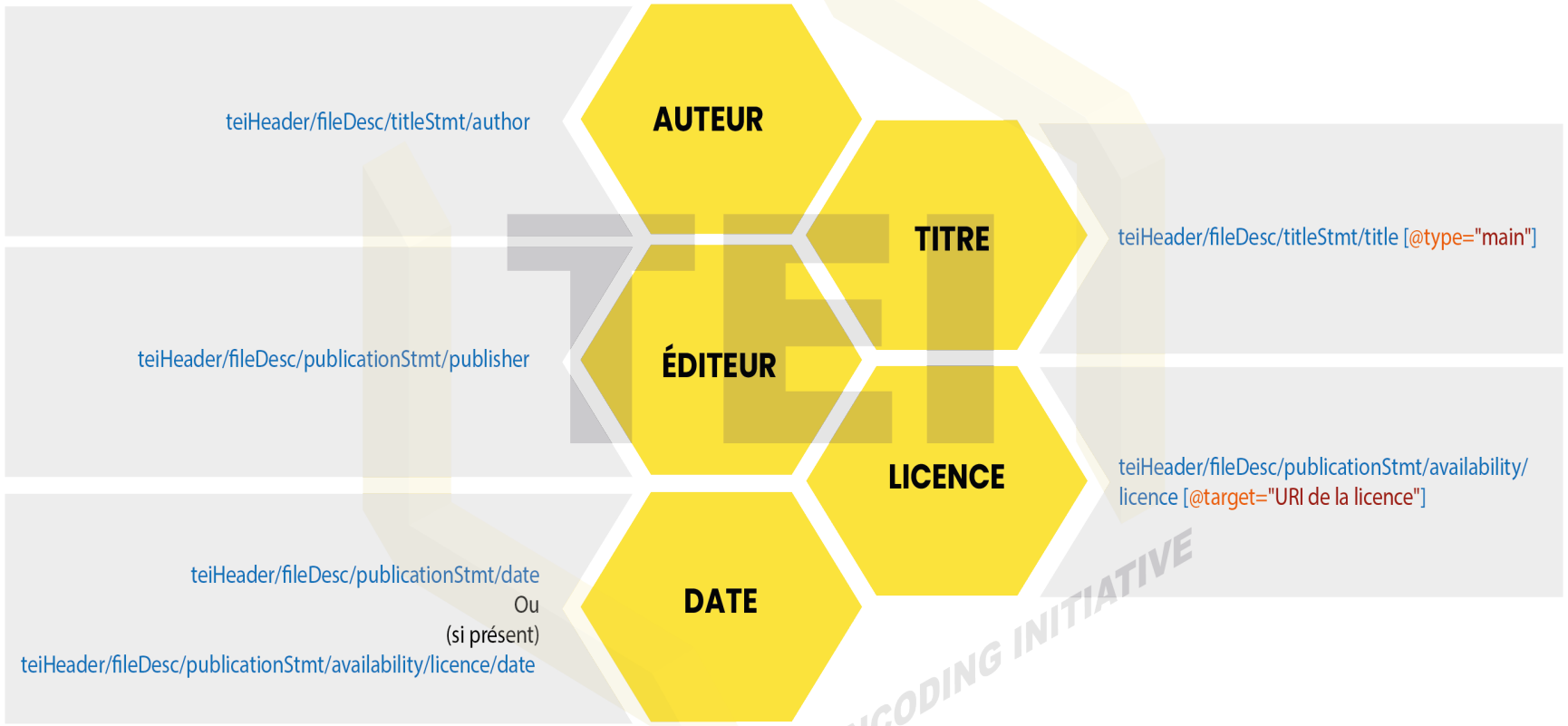
API Nakala

Point d'accès SPARQL Nakala

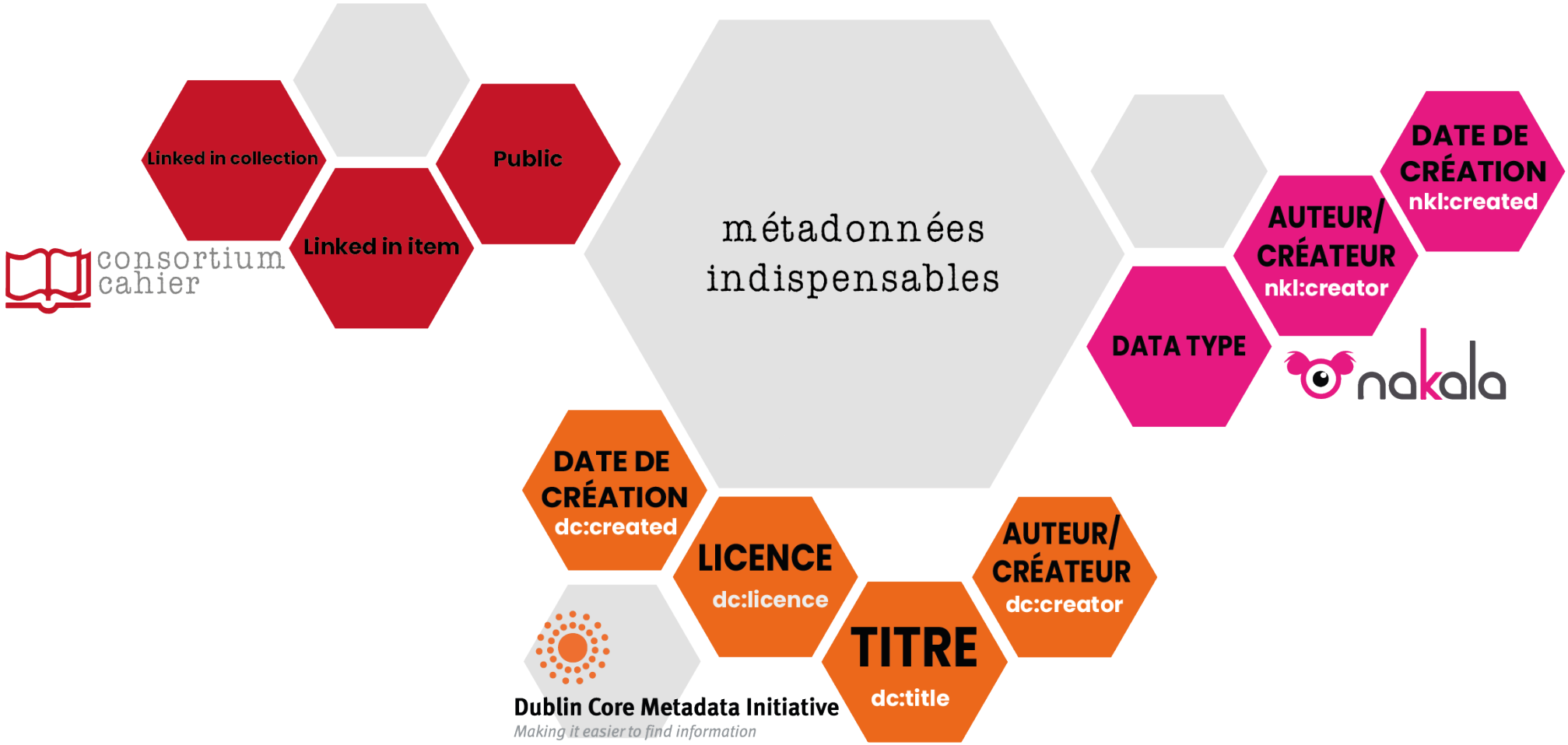


L'outil mynkl

L'outil mynkl organise le mapping de vos données en s'appuyant sur la sélection de balises TEI suivantes :

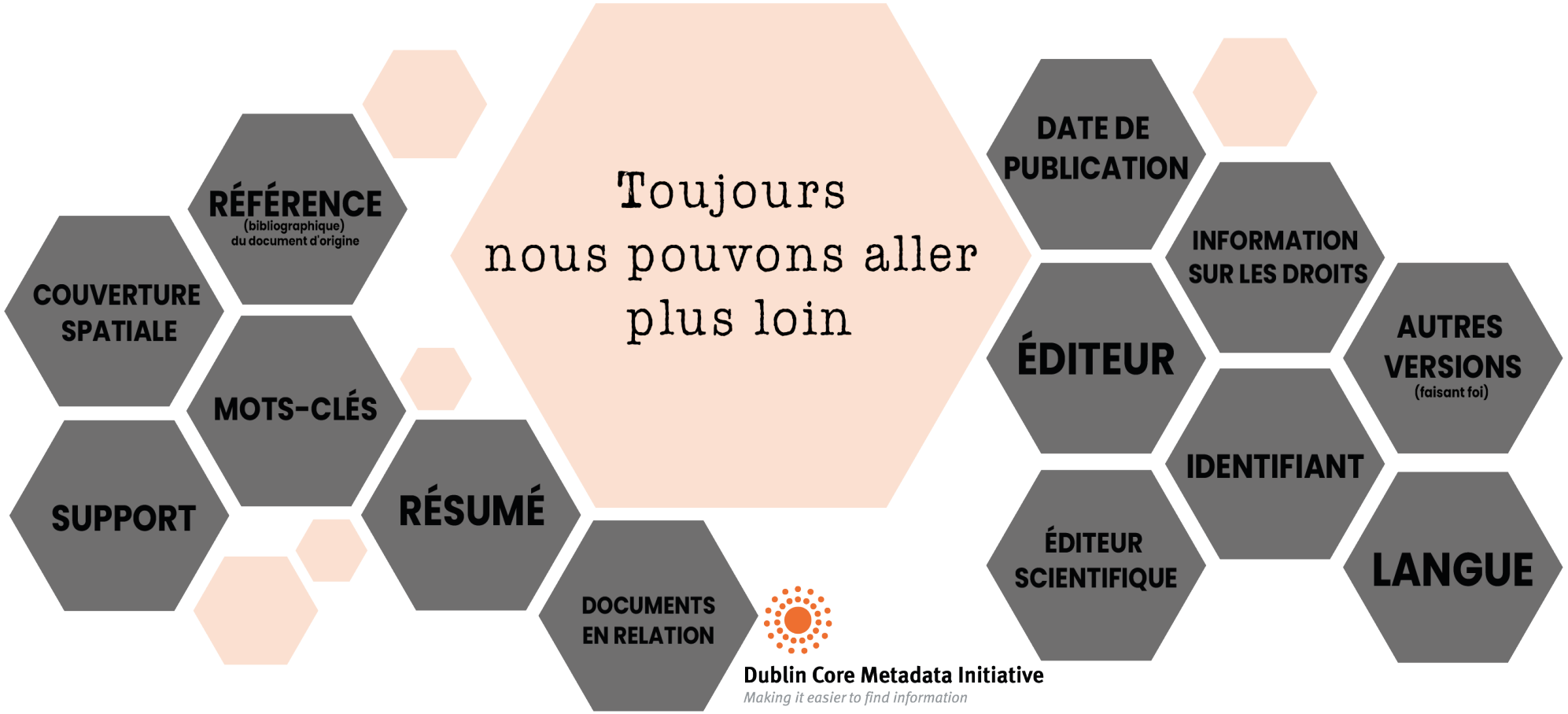


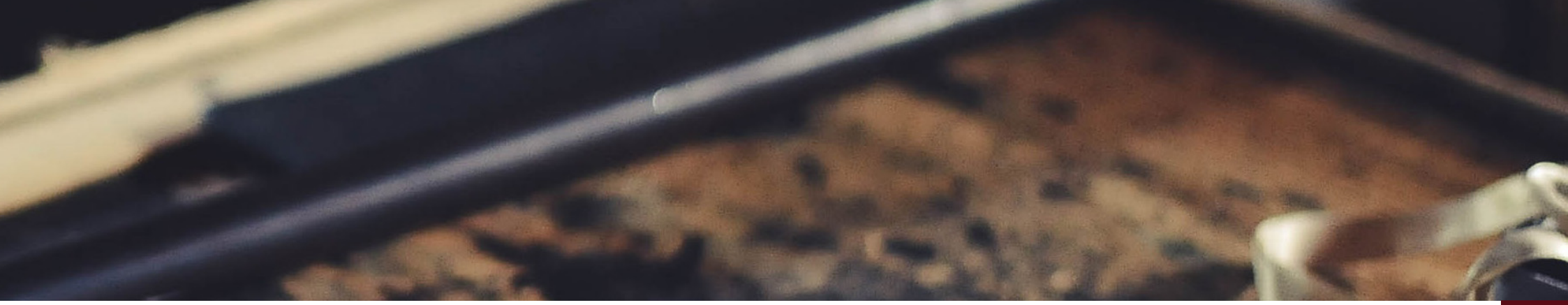
L'outil mynkl organise le mapping de vos données en s'appuyant sur la sélection des champs DC suivants :



L'outil mynkl

Si vous utilisez un tableur et les métadonnées suivantes, l'outil mynkl organise le mapping de vos données





Si vous utilisez un tableur, voici une liste de champs de métadonnées que nous vous suggérons

Linked in collection
Public
Linked in item
Format dc-rdf:format
DataType
Titre dc-rdf:title
Créateur nkl:creator
Créateur dc-rdf:creator
Couverture spatiale dc-rdf:spatial
Date de création dc-rdf:created
Date de création nkl:created
Licence : dc-rdf-license
Langue dc-rdf:language
Sujet dc-rdf:subject
Description dc-rdf:description
Contributeur dc-rdf:contributor
Éditeur dc-rdf:publisher
Ayants droit dc-rdf:rightsHolder
Relation dc-rdf:relation
Résumé dc-rdf:abstract
Date de disponibilité dc-rdf:available
Date de modification dc-rdf:modified
Est une version de dc-rdf:isVersionOf
Support dc-rdf:medium
Référence bibliographique dc-rdf:bibliographicCitation

Les métadonnées

À propos de quelques métadonnées

Titre dc-rdf:title

Titre dc-rdf:title : sert à déterminer le nom donné à la ressource
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.]

Langue dc-rdf:language

Ce champ indique la (ou les) langue.s de la ressource. Il existe différentes représentations des langues possibles reconnues par Nakala. On peut retrouver cette liste sur le site de l'API de Nakala à partir du lien lié au QR code Langues Nakala.

Langues Nakala

Public

Il est nécessaire de prêter une attention particulière à ce point, car une fois que les données sont publiques dans Nakala, elles ne peuvent plus être retirées du serveur, à moins d'une communication directe entre les chercheurs et les administrateurs de Nakala.

Lorsque le dépôt est effectué et que la donnée est publique, elle devient citable, c'est pour cela qu'elle ne peut plus être supprimée.

Date de création nkl:created - Date de création dc-rdf:created

Ce champ détermine la date de création de la ressource. La pratique recommandée est de décrire la date, la date/heure ou la période. Il s'agit de la date de création du document source (appelée date première dans le catalogue OAI du consortium CAHIER). La date doit être inscrite suivant la norme ISO 8601, c'est-à-dire AAAA-MM-JJ.

Les métadonnées

DataType

Ce champ sert à déterminer le type de données que le chercheur dépose dans les serveurs de Nakala. Pour bien renseigner cet espace il faut coller dans le champ dédié le lien URI correspondant au type de donnée parmi la liste suivante :

Image	"http://purl.org/coar/resource_type/c_c513"
Vidéo	"http://purl.org/coar/resource_type/c_12ce"
Son	"http://purl.org/coar/resource_type/c_18cc"
Articles de journaux	"http://purl.org/coar/resource_type/c_6501"
Poster dans une conférence	"http://purl.org/coar/resource_type/c_6670"
Objet présenté dans une conférence	"http://purl.org/coar/resource_type/c_c94f"
Objet d'apprentissage	"http://purl.org/coar/resource_type/c_e059"
Ouvrage	"http://purl.org/coar/resource_type/c_2f33"
Carte ou une carte géographique	"http://purl.org/coar/resource_type/c_12cd"
Jeu ou une série de données	"http://purl.org/coar/resource_type/c_ddb1"
Logiciel ou un développement informatique	"http://purl.org/coar/resource_type/c_5ce6"
Ressource qui n'est pas incluse parmi la liste de controlled vocabulaires for repositories (COAR)	"http://purl.org/coar/resource_type/c_1843"

Matériel d'archive	"http://purl.org/library/ArchiveMaterial"
Collection [URI DublinCore]	"http://purl.org/ontology/bibo/Collection"
Bibliographie	"http://purl.org/coar/resource_type/c_86bc"
Séries [URI DublinCore]	"http://purl.org/ontology/bibo/Series"
Note de lecture, une critique de livre ou une recension d'ouvrage	"http://purl.org/coar/resource_type/c_ba08"
Manuscrit	"http://purl.org/coar/resource_type/c_0040"
Lettre	"http://purl.org/coar/resource_type/c_0857"
Rapport	"http://purl.org/coar/resource_type/c_93fc"
Periodique [c'est un concept rendu obsolète par la COAR mais avec une URI existante et utilisé par Nakala pour l'instant	"http://purl.org/coar/resource_type/c_2659"
Prépublication	"http://purl.org/coar/resource_type/c_816b"
Synthèse, un article de synthèse ou une recension	"http://purl.org/coar/resource_type/c_efa0"
Partition de musique	"http://purl.org/coar/resource_type/c_18cw"
Ensemble des données collectées enquête en considérant les complétions effectuées par les participants.	"https://w3id.org/survey-ontology#SurveyDataSet"

Les métadonnées

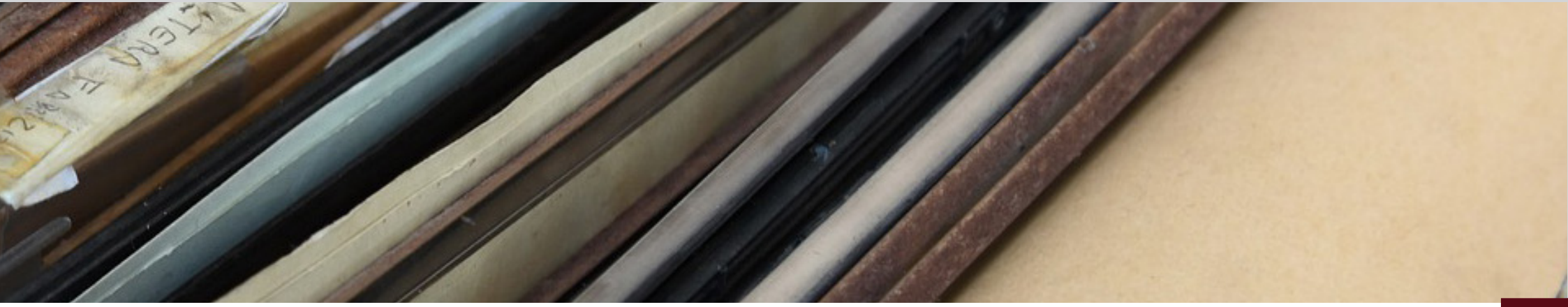
Texte simple	“http://purl.org/coar/resource_type/c_18cf”
Thèse, une mémoire de thèse ou un rapport de thèse.	“http://purl.org/coar/resource_type/c_46ec”
Site web	“http://purl.org/coar/resource_type/c_7ad9”
Data paper ou article sur les données	“http://purl.org/coar/resource_type/c_beb9”
Ressource interactive	“http://purl.org/coar/resource_type/c_e9a0”

Créateur dc-rdf:creator - Créateur nkl:creator

Ce champ détermine l’entité principalement responsable de la fabrication de la ressource. Creator peut comprendre une personne, une organisation ou un service. Les noms des auteurs dans cet espace s’écrivent [prénom][virgule][espace][nom]

Exemple 1 : Jean-Paul, Sartre
Exemple 2 : Emile, Zola
Exemple 3 : Julio, Cortazar

Liste DataTypes API Nakala



Référence bibliographique dc-rdf:bibliographicCitation

Ce champ détermine s'il s'agit d'un livre, d'un article, ou d'une autre ressource documentaire.

Couverture spatiale dc-rdf:spatial

Ce champ détermine les caractéristiques spatiales de la ressource.

Ayants droit dc-rdf:rightsHolder

Ce champ détermine la personne ou l'organisation qui possède ou gère les droits sur la ressource. La pratique recommandée est de faire référence au détenteur des droits avec une URI. Si cela n'est pas possible ou faisable, une valeur littérale qui identifie le titulaire des droits peut être fournie.

Les métadonnées

Licence : dc-rdf-license

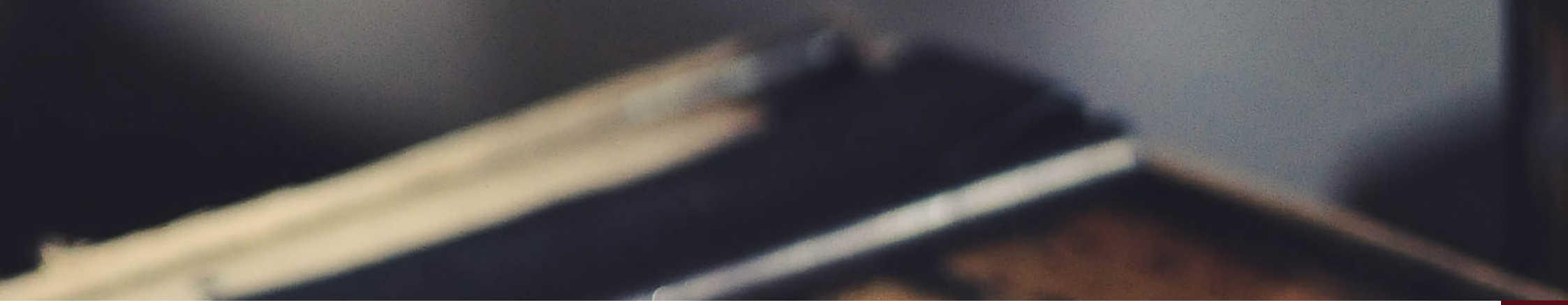
Ce champ détermine le document légal ou le type d'autorisation délivrée avec la ressource. La pratique recommandée est d'identifier la licence avec une URI. Si cela n'est pas possible ou faisable, une valeur littérale qui identifie la licence peut être fournie. Les différentes licences possibles peuvent se retrouver sur le site de l'API de Nakala en suivant le lien lié au QR code Licences Nakala.

Licences Nakala

Source dc-rdf:source

Ce champ indique la référence ou côte de la ressource connexe à partir de laquelle la ressource numérique décrite est dérivée. La pratique recommandée est d'identifier la ressource connexe au moyen d'une chaîne conforme à un système d'identification formel. On mentionnera ici les données bibliographiques du document source (date, lieu et année de publication de la source, côte du document dans l'institution).

Exemple 1 : NAF 10266
Exemple 2 : Cote : NAF 10266



Éditeur dc-rdf:publisher

Ce champ détermine l'entité responsable de la mise à disposition de la ressource.

Exemple 1 : Projet Région n°12345 | Consortium CAHIER TGIR Huma-Num
Exemple 2 : Projet FLG – AAP MSH Centre Sud

Contributeur dc-rdf:contributor

Ce champ détermine l'entité responsable des contributions à la ressource. Les directives relatives à l'utilisation de noms de personnes ou d'organisations en tant que créateurs s'appliquent également aux contributeurs. En général, le nom d'un contributeur doit être utilisé pour indiquer l'entité.

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)

Les métadonnées

Droits dc-rdf:rights

Ce champ détermine les informations sur les droits détenus dans et sur la ressource. En général, les informations comprennent une déclaration sur les divers droits de propriété associés à la ressource, y compris les droits de propriété intellectuelle.

Exemple 1 : Fonds Jean-Charles Carpo – Bibliothèque Jacques Tecuod
Exemple 2 : Archives familiales Jean-Sol Ertrap

Relation dc-rdf:relation

Ce champ désigne une ressource connexe. La pratique recommandée est d'identifier la ressource connexe au moyen d'une URI. Si cela n'est pas possible ou faisable, une chaîne conforme à un système d'identification formel peut être fournie.

Exemple 1 : Le rose et le jaune – Manuscrit 1 | Le rose et le jaune – Manuscrit 3 | Le rose et le jaune – Carnet 1
Exemple 2 : NAF 10266 | NAF 10267 | NAF 10268

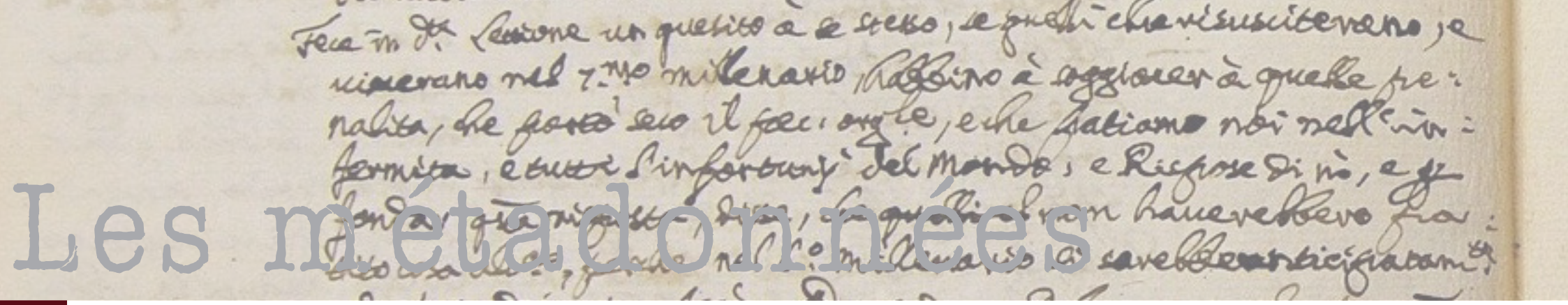
Format dc-rdf:format

Ce champ détermine le format du fichier. La pratique recommandée est d'utiliser un vocabulaire contrôlé lorsqu'il est disponible. Par exemple, pour les formats de fichiers, on peut utiliser la liste des Internet Media Types (MIME).

Résumé dc-rdf:abstract

Ce champ comporte quelques lignes et informations à propos de la ressource. Elles sont destinées au public.

Exemple 1 : Le roman parle de ....
Exemple 2 : Ce document est le premier de...

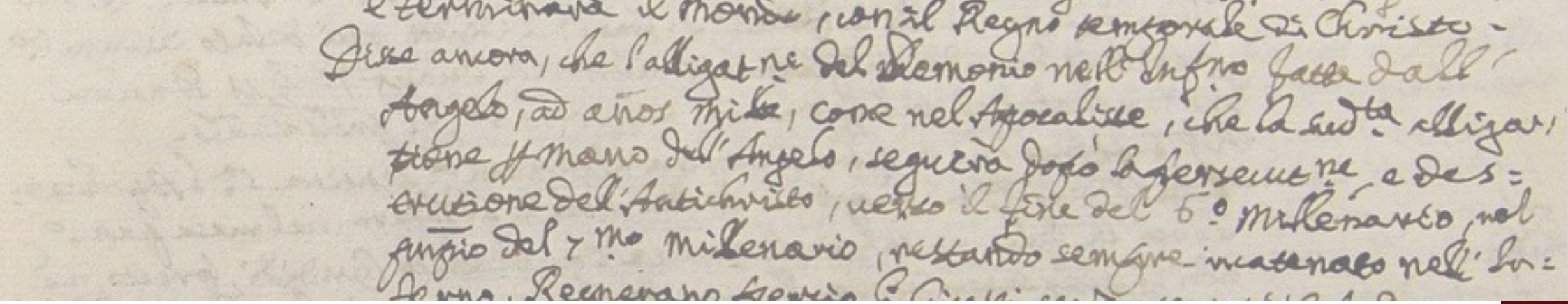


Date de disponibilité dc-rdf:available

Ce champ détermine la date à laquelle la ressource est devenue ou deviendra disponible.
Comme recommandé pour la propriété Date, dont Available est une sous-propriété, la date doit être inscrite suivant la norme ISO 8601, c'est-à-dire AAAA-MM-JJ. Il est donc possible de mettre des données sous embargo.

Date de modification dc-rdf:modified

Ce champ détermine la date à laquelle la ressource a été modifiée. La pratique recommandée est de donner la date, la date/heure ou la période. La date doit être inscrite suivant la norme ISO 8601, c'est-à-dire AAAA-MM-JJ.



Est une version de dc-rdf:isVersionOf

Ce champ indique la ressource connexe dont la ressource décrite est une version, une édition ou une adaptation. Les changements de version impliquent des modifications substantielles du contenu plutôt que des différences de format. Cette propriété est destinée à être utilisée avec des valeurs non-littérales. Cette propriété est une propriété inverse de Has Version. 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 : Le rose et le jaune – Manuscrit 2

Support dc-rdf:medium

Ce champ indique le support matériel ou physique de la ressource.

FAIRiser vos données : mémorandum