Procédure détaillée de création de fiches de métadonnées

Parc national des Ecrins

Juillet 2020

laboratoire edytem environnements dynamiques territoires montagnes

INTRODUCTION

Dans une tendance générale en faveur de l'Open Data, l'interconnexion des données est devenue un enjeu fondamental. L'interopérabilité s'effectue notamment par la création de métadonnées qui répondent aux enjeux pour la réutilisation et la valorisation des données et qui permettent la création de nouvelles connaissances. Les protocoles de production de ces métadonnées nécessitent néanmoins l'utilisation de formats d'échange définis et de standards dans le but de maintenir l'interopérabilité des données entre elles et d'assurer leur pérennisation.

Le protocole de production détaillé dans cette fiche utilise logiciel R version 3.6.2 (2019-12-12). La procédure s'appuie sur la Package R ***metadataReasy***, créée en 2018 par Christine Plumejeaud (LIENSS) en lien avec Juliette Fable (OSU OREME. Le protocole est disponible sur Github avec le lien suivant : [https://github.com/cplumejeaud/metadataReasy]. La procédure utilise le langage de programmation R avec un script complet permettant la production automatisée de fiches de métadonnées. Le script extrait toutes les métadonnées d'un fichier Excel et les formate selon la norme ISO19115 via la librairie Geomata. Le script est basé sur l'exemple Geomata Isometdata (voir la documentation Geomata) et Geonapi (voir la documentation Geonapi). La procédure détaillée est complétée par la procédure d'import des fiches nouvellement créées dans le catalogue du Parc National des Écrins.

Ce script utilise sur les packages et librairies R suivants :

- *XML*
- *geometa*
- *uuid*
- *rgeos*
- *gdata*

L'exécution du script nécessite, par ailleurs, une puissance minimale d'ordinateur. En effet, le chargement de catalogues ainsi que la structure du script tendent à être difficilement supportables pour un ordinateur de puissance faible.

PROTOCOLE DE PRODUCTION DES FICHES DE MÉTADONNÉES

JEUX DE DONNÉES

Une bonne connaissance des jeux de données étudiés représente le point primordial à considérer en amont de la production des fiches de métadonnées afin d'assurer l'exactitude des informations au sein des fiches.

COMPLÉTION DU FICHIER

La procédure s'appuie sur un fichier modèle Excel® (*.xslx*). Ce dernier, pré-rempli, répond à la norme ISO 19115. Ce fichier dispose de plusieurs autres onglets aidant à la complétion de la fiche modèle.

L'onglet '*metadata*' est l'onglet à compléter. **Chaque ligne correspond à un jeu de données qui sera à l'origine d'une fiche de métadonnées**.
L'onglet '*guidelines*' est l'onglet guide pour la saisie de chaque colonne. Il est nécessaire de suivre ces instructions pour la validité de la fiche de métadonnées sortante.
Certains champs, tel que online_resource, restent néanmoins facultatifs.

Les indications de saisie sont les suivantes :
Nb : Les champs annotés en **gras** ci-dessous sont les champs de saisie obligatoires.

À noter : Le texte doit être encodé en UTF-8 afin d'éviter les problèmes d'encodage.
Les accents peuvent être ajoutés au texte du fichier modèle.

- ✔ **resource_identifier**, parent_identifier et **status** ne sont pas à saisir immédiatement. Ces champs seront saisis automatiquement par le script R.

- ✔ **resource_identifier** : Identifiant unique de la donnée (comprenant, par exemple, le numéro de SIRET de l'institution)

✔ parent_identifier : Identifiant de la série « parent » si le jeu de données fait partie d'une série plus large ayant fait l'objet elle-même d'une fiche de métadonnées.

✔ **status** : Statut de la fiche(en cours de complétion ou terminée).
Le vocabulaire à utiliser est 'terminée' (completed) ou 'en cours de complétion' (being completed).

À noter : Dans le but d'éviter un écrasement automatique des fiches sortantes par la dernière nouvellement créée, il est conseillé de donner un nom explicite à chaque jeu de données dans la colonne **resource_identifier**. Ce champs peut être modifié par la suite automatiquement, en fonction de l'interface d'accueil de la fiche, afin de saisir l'identifiant réel de la donnée.

✔ **title** : Un titre explicite comprenant si possible des **mots-clés**, un **lieu** associé ainsi qu'une **date** (à minima l'année). Les acronymes sont déconseillés.

Type de donnée attendue : Texte
Exemple : Mesure Colibri SN0641 aout 2015 - serie 1 ZATU (Zone Atelier Territoires Uranifères (LTER-France)

✔ **abstract** : Un bref résumé présentant la donnée (contexte et enjeux, objectifs…). Ce résumé doit permettre une bonne description de la donnée. Un URL peut également être ajoutée pour plus d'informations.

Type de donnée attendue : Texte
Exemple : Emplacements des prélèvements d'eau faits sur et autour du site de Rophin avec la nature du prélèvement (surface, porale…). Une partie de ces prélèvements a été réalisée dans le cadre du projet TREMPLIN. Contributeurs : Subatech , IPHC,LPC, IRSN, LMGE, CENBG

✔ resource_type : Il s'agit d'indiquer le type de donnée (séries de données géographiques, ensemble de données géographiques, services de données géographiques).

Type de donnée attendue : Texte ou liste (vocabulaire contrôlé). Le vocabulaire à utiliser est le suivant : 'dataset' ou 'series'.
Exemple : dataset, series

✔ spatialRepresentationType : Type de représentation.

Type de donnée attendue : Texte ou liste (vocabulaire contrôlé) Le vocabulaire à utiliser est le suivant : 'vector' ; 'grid' ; 'textTable'.
Exemple : vector, grid

✔ resource_language : Préciser la langue dans laquelle les données sont décrites

Type de donnée attendue : Texte ou liste (vocabulaire contrôlé). Le vocabulaire à utiliser correspond à la liste des codes ISO639.
Exemple : fre, eng

✔ creation_date : Date de création de la ressource, la ressource correspondant à la fiche de métadonnées.

Type de donnée attendue : Date
Exemple : 01/02/2019

À noter : Il est nécessaire que l'ensemble des dates décrites dans le fichier soient au format Jour/Mois/Année ou sous le format Année-Mois-Jour pour le bon fonctionnement du script R.
Si la date décrite n'est pas sous l'un de ces formats, aucune fiche XML ne pourra être créée.

✔ publish_date : Date de publication de la ressource. Cette date correspond à la date à laquelle la ressource a été mise en ligne suite à sa création.

Type de donnée attendue : Date
Exemple : 01/02/2019

✔ update_date : Date de la dernière mise à jour en ligne de la ressource. Chaque mise à jour doit être notifiée.

Type de donnée attendue : Date
Exemple : 05/02/2019
Use case : Les jeux de données récoltés sur le long terme nécessite dans certains cas des mises à jour en continu. Il est alors nécessaire d'indiquer chaque mise à jour sur la ressource.

✔ resource_format : Format de la ressource

Type de donnée attendue : Texte (vocabulaire contrôlé). Le vocabulaire à utiliser est le suivant : 'tableDigital' ; 'mapDigital' ; 'imageDigital' ou 'text'.
Exemple : tableDigital, mapDigital

✔ update_frequency : Fréquence des mises à jour.

Type de donné attendue : Texte ou liste (vocabulaire contrôlé). Le vocabulaire à utiliser est le suivant : 'continual' ou 'notPlanned'.
Exemple : continual / notPlanned

✔ temporal_extent_name : étendue temporelle correspond à une période de temps couverte par le contenu de la ressource.

Type de donnée attendue : date ou intervalle de date ou mélange dates et intervalles (texte)
Exemple 1 : 2009-05-01
Exemple 2 : 2009-05-01 au 2010-04-01

✔ start_date : Date de début de la ressource

Type de donnée attendue : Date
Exemple : 2009-05-01
Use case : En prenant comme exemple les carottes sédimentaires du lac de la Muzelle dans le Parc National des Écrins, la période couverte par la fiche correspond à la période de temps étudiée à travers les sédiments. Ces derniers couvrent la période de l'an 303 à l'an 2012. La date de début de la ressource est 0303-01-01.

✔ end_date : Date de fin de la ressource

Type de donnée attendue : Date
Exemple : 2009-05-01
Use case : Avec l'exemple des carottes sédimentaires du lac de la Muzelle, la date de fin est la suivante : 2012-01-01.

- ✔ **spatial_extent_name** : Nom de la zone géographique considérée. Les catalogues internationaux tels que Geonames (https://www.geonames.org/) ou LocationIQ (https://locationiq.com/) sont à favoriser pour la complétion de ce champ.

  Type de donnée attendue : Texte
  Exemple : Lachaux, Bessines

- ✔ **geom** : Coordonnées géographiques de l'emprise spatiale de la ressource. Les coordonnées sont a ajouter également dans l'onglet 'boundingbox'.

  Texte de donnée attendue : Texte
  Exemple :POLYGON((6.097751355112142 44.952416757421936,6.09955379957015 44.95086781949248,6.099639630258626 44.94898474061112,6.098395085275716 44.948832876718946,6.094833111703939 44.949987032222325,6.095262265146322 44.95108042129124,6.095777249277181 44.95144487968543,6.095948910654134 44.9519308206109,6.097751355112142 44.952416757421936))

À noter : Se connecter à https://arthur-e.github.io/Wicket/sandbox-gmaps3.html et dessiner l'emprise de des données, puis copier le Well-Known Text (WKT)

- ✔ **reference_system** : Référentiel de coordonnées.

  Type de donnée attendue : texte ou liste. Les référentiels les plus courants sont les suivants : 4326 (WGS 84) , 2154 (RGF93 / Lambert-93), 3946 (RGF93 / CC46), 27572 (NTF (Paris) / Lambert zone II)
  Exemple : 4326

À noter : L'OGC fournit un espace de noms pour référencer les systèmes de référence, par exemple, http://www.opengis.net/def/crs/EPSG/0/4258 est la référence du système ETRS89 dans le registre EPSG.

- ✔ **topic_categories** : définir la thématique ou les thématiques. Le vocabulaire est contrôlé selon la Norme ISO (IsoTopicCategory).

  Type de donnée attendue : Texte
  Exemple :environment---http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/environment

À noter : Se connecter à : http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/. Cliquer sur une étiquette et copier la fin de l'url : http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/environment. Séparer d'une virgule, si plusieurs thématiques

À noter : Le choix de la langue est laissé libre au producteur des données. Néanmoins, les thèmes INSPIRE étant associés à des codes, le champ topic_categories doit être complété avec la version anglaise des thèmes.

- ✔ **inspire_themes** : définir le thème INSPIRE

  Type de donnée attendue : Texte
  Exemple : ressources minérales---http://inspire.ec.europa.eu/theme/mr, lieux de production et sites industriels---http://inspire.ec.europa.eu/theme/pf, installations de suivi environnemental---http://inspire.ec.europa.eu/theme/ef

À noter : Se connecter à : http://inspire.ec.europa.eu/theme/
Cliquer sur une étiquette et copier le nom du thème, suivi par --- et de l'url, aucun espace n'est nécessaire. Vous ne devez choisir qu'un seul thème INSPIRE

- ✔ **gemet_keywords** : définir les mots-clés GEMET (General Multilingual Environmental Thesaurus)

  Type de donnée attendue : Texte
  Exemple : industrie minérale---https://www.eionet.europa.eu/gemet/fr/concept/5268, sol contaminé---https://www.eionet.europa.eu/gemet/fr/concept/1751, analyse de l'eau---http://www.eionet.europa.eu/gemet/concept/9147

À noter : Se connecter à : https://www.eionet.europa.eu/gemet/fr/themes/
Choisir un thème, puis un mot-clé et copier le nom du concept, suivi par --- et de l'url. Si plusieurs mots-clés sont nécessaires pour décrire la donnée, il est nécessaire de les séparer d'une virgule.

ATTENTION : La séparation entre 2 mots-clés ne nécessite qu'une simple virgule, sans espace ajouté.

- ✔ **other_keywords** : Autres mots-clés issus de vocabulaires ou thesaurus thematiques, métier ou nomenclatures tels que EnvThes, GCMD, etc.

  Type de donnée attendue : texte
  Exemple : levés topographique; GPS, etc.

À noter : D'autres mots-clés peuvent être ajoutés, séparés par un « ; ».
Le site européen Fairsharing propose un moteur de recherche de standards et de vocabulaires. (https://fairsharing.org/).
Le format d'insertion est laissé libre à l'utilisateur. Il ainsi possible d'utiliser le même format d'insertion que celui des mots clés GEMET, soit : mot-clé---lien.

- ✔ md_contact : compléter les adresses e-mail suivantes : l'auteur de la ressource, suivie de la principale partie chargée de recueillir des informations et de mener les recherches (maître d'œuvre) et enfin la personne qui peut être contactée pour s'informer sur la ressource. L'onglet 'contact' est également à compléter avec les informations concernant les auteurs et points de contact.

  Type de donnée attendue : Texte
  Exemple : author=xxxxx@univ-lr.fr;principalInvestigator=xxxx@univ-lr.fr; pointOfContact=xxxx@univ-lr.fr

À noter : L'adresse générique de l'institution ou du gestionnaire de la donnée du laboratoire est à privilégier pour la complétion du champ md_contact, les adresses personnelles pouvant évoluer au cours du temps.

- ✔ lineage : décrire la généalogie, l'historique des données : comment elle a été collectée, créée, transformée… (matériels, méthodes, protocoles…). La complétion de ce champ doit permettre une meilleure compréhension de la donnée afin de garantir au maximum une ré-utilisation adaptée de celle-ci.

  Type de donnée attendue : Texte
  Exemple : Les données contiennent une position GPS et une localisation site (dans la prairie…). Pour les carottes, la profondeur et le nombre de tranches sont précisés. Les différents types d'analyse sont indiqués : Gamma, Perte au feu , Granulometrie, Mineralogie (DRX), Extraction sequentielle, Extraction selective, ICP (apres mineralisation), XRF-portable, Echange isotopique, EXAFS, XAS, SEM-EDX, EPMA, Datation carbone 14. Dans le cadre des carottes sédimentaires, sont précisés également ici les nombres de lamines prélevées et leur résolution, le paramètrage particulier des machines non inclus dans le fichier de sortie (exemple "utilisation des ultrasons lors de l'analyse (low, medium ou hight"pour la granulométrie),

✔ use_condition :Préciser les conditions d'utilisation, les droits associés à la ressource (copyright, etc.).

Type de donnée attendue : Texte
Exemple : This work is licensed under a Creative Commons Attribution 4.0 License (CC BY SA 4.0). Pour en savoir plus : https://creativecommons.org/licenses/by-sa/4.0/).

À noter : Si celles-ci ne sont pas précisées, elles seront, par défaut, définies sous licence Creative Commons Attribution 4.0 (CC BY SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/). L'œuvre peut être librement utilisée, à la condition de l'attribuer à l'auteur en citant son nom. Cela ne signifie pas que l'auteur est en accord avec l'utilisation qui est fait de ses œuvres.

✔ online_resource : mettre le lien vers un site présentant le jeu de donnée. Par défaut, s'il n'est pas précisé, le site du laboratoire sera indiqué.

Type de donnée attendue : URL
Exemple : https://lienss.univ-larochelle.fr/

✔ thumbnail_url : mettre le lien vers une illustration de type imagette, s'il existe. Il est possible dans ce champs d'intégrer le lien vers une cartographie des données par exemple.

Type de donnée attendue : URL
Exemple : Les données de bathymétrie d'un lac peuvent être modélisées. Ce champ peut donc permettre d'intégrer le lien vers cette modélisation.

✔ wms_resource : indiquer l'adresse du flux de données WMS

Type de donnée attendue : URL

Fichier d'origine => Extraction en CSV

Le fichier doit être homogène dans sa saisie. Une ligne doit correspondre à un seul jeu de données. De plus, la procédure via le script R n'est capable de supporter que des fichiers au format ***.csv***.

Une extraction des onglets d'intérêt à partir du fichier modèle est ainsi donc indispensable.L'onglet '*metadata*' et l'onglet '*contacts*', dûment complétés, sont à extraire du fichier au format ***.csv***. Une fois extraits, ces fichiers pourront être intégrés au script R de création des fiches.

Exécution du script R

Dans le script, 3 variables de configuration doivent être adaptées ; elles correspondent :

- au répertoire d'entrée (répertoire au sein duquel le script va récupérer les fichiers à traiter),
- au répertoire de sortie (répertoire au sein duquel le script va enregistrer les fiches de métadonnées),
- au préfixe défini pour les fiches. Le préfixe sera utilisé par le script pour définir l'identifiant de la ressource.

Après création des répertoires, le chemin associé doit être intégré au script de la manière suivante : Les sections à personnaliser sont indiquées en bleu.

Répertoire d'entrée où se situe les fichiers CSV de métadonnées et contacts :

metadatadir <- "D:/Travail/OwnCloud/Zone Atelier PVS/Qrcode/Qrcode_3/Metadata/Cours_atelier_R_MD/ZATU/"

Répertoire de sortie où seront exportées les fichiers XML de métadonnées

exportxml_dir <- "D:/Travail/OwnCloud/Zone Atelier PVS/QRcode/QRcode_3/Metadata/Cours_atelier_R_MD/Export_XML/"

Prefix des fiches de métadonnées

prefix <- "ZATU"

À noter : Le chemin du répertoire doit contenir uniquement les symboles suivants : **« / », « : »**.
Le chemin doit obligatoirement se terminer par le symbole **« / »**.
Il est également nécessaire d'inclure le nom complet des fichiers csv 'contacts' et 'metadata' au sein du script.

```
metadata <- read.csv(paste0(metadatadir,"metadata_ISO19115_pour_scriptR-ZATU-V3_3_metadata.csv"), sep=";", dec=".", stringsAsFactors=FALSE)
contacts <- read.csv(paste0(metadatadir,"metadata_ISO19115_pour_scriptR-ZATU-V3_3_contacts.csv"), sep=";", dec=",", stringsAsFactors=FALSE)
```

Le script ainsi personnalisé peut être exécuté.

Fiches de métadonnées (XML)

Les fiches XML sont enregistrées dans le répertoire de sortie indiqué. Elles sont au format XML, particulièrement adapté pour l'échange de données numériques. Le nom des fiches sortantes correspond au nom indiqué dans la colonne **resource_identifier**.