

Un workflow R pour une gestion simple et durable des données de la recherche

Wilfried Heintz¹, Julien Barde², Emmanuel Blondel³

Résumé. A l'heure de l'Open Science, même si de nombreuses initiatives locales, institutionnelles et nationales prouvent que le partage des données scientifiques est désormais une nécessité, force est de constater que nous nous heurtons toujours aux mêmes difficultés de mise en œuvre. En effet, un des fondements de la science ouverte et du principe de données FAIR⁴ (Faciles à trouver, accessibles, interopérables et réutilisables) est la fourniture de métadonnées : ce sont des informations contextuelles permettant de découvrir un jeu de données et de le comprendre pour le réutiliser. Or il existe toujours des freins à la création de ces métadonnées, principalement techniques, faute d'outils simples et faciles à prendre en main par des non spécialistes, et ce, malgré les considérables évolutions des logiciels les plus connus, tels que Geonetwork.

Un workflow a été implémenté en langage R - largement utilisé par la communauté scientifique -, permettant à partir d'un tableur partagé, éditable dans un environnement collaboratif, de générer des métadonnées conforme à la norme ISO/OGC 19115 (également d'autres standards comme l'EML), de les publier dans un catalogue de métadonnées et/ou dans un ou plusieurs entrepôts de données mais également de créer un flux de données. Nous nous proposons de détailler les mécanismes et les fonctionnalités de ce workflow.

Mots clés : R, gestion de données, FAIR, Infrastructure de données spatiales, Base de données, Workflow, Plan de gestion de données, Open Science, Entrepôt de données

Abréviations et acronymes

API : Application Programming Interface (Interface de programmation applicative en français)

CSW: Catalogue Service for the Web

FAIR : Findable, Accessible, Interoperable, Reusable

INRA : Institut National de la Recherche Agronomique

IRD : Institut pour la Recherche et le Développement

IDG (ou IDS) : Infrastructure de Données Géographiques (ou Spatiales)

JSON : Javascript Object Notation

OGC : Open Geospatial Consortium

OS : Operating System (Système d'exploitation en français)

PGD : Plan de gestion de données

W3C : World Wide Web Consortium

WFS : Web Feature Service

WMS: Web Map Service

¹ UMR Dynafor, INRA, F-31327 Castanet-Tolosan, France ; wilfried.heintz@inra.fr

² UMR MARBEC, IRD, F-34203 Sète, France

³ Consultant indépendant, F-14000 Caen, France

⁴ <https://www6.inra.fr/datapartage/Produire-des-donnees-FAIR>

Introduction

De nombreux instituts de recherche se sont engagés dans le partage et la diffusion de leurs données scientifiques, comme l'atteste par exemple la charte de libre accès aux publications et données scientifiques de l'INRA publiée en 2016⁵. Bien avant cet engagement, d'importants efforts ont été portés par les géomaticiens sur la structuration des données scientifiques, conformément aux recommandations des standards internationaux (OGC, W3C) et de la directive européenne INSPIRE⁶. La mise en oeuvre d'IDG à plusieurs échelles (Unités de recherches, départements scientifiques, instituts) a contribué à une meilleure gestion des données. Les outils de catalogage, tels que Geonetwork, sont des composants essentiels de ces IDG qui ont été considérablement améliorés depuis leur version initiale (en particulier l'ergonomie des interfaces graphiques).

Malgré ces évolutions, force est de constater que les catalogues de métadonnées restent en général peu remplis au regard des données existantes ce qui limite les possibilités de réutilisation des jeux de données. A l'échelle d'une unité de recherche, une explication possible est que les personnes les plus légitimes pour créer ces métadonnées (collecteur de données, responsables de projets) ont trop peu de temps à consacrer à la prise en main d'un outil informatique spécialisé, sachant que la rédaction des métadonnées elle-même est une activité chronophage, notamment si elle n'est pas intégrée dès le début de la collecte ou de la production de données.

Partant de ce constat, dans le cadre d'un projet H2020, une chaîne de traitements (ou workflow) codée en langage R a été mise en place pour automatiser la création et la mise à jour de fiches de métadonnées conformes aux standards de l'OGC à partir d'un tableur partagé. L'objectif initial est de mettre à disposition des personnes ressources pour le catalogage (qui ne sont pas les gestionnaires de données ou géomaticiens) des outils simples et couramment utilisés dans le cadre de leurs activités.

Nous proposons de détailler le fonctionnement de cet outil, de l'environnement partagé qu'il propose et des possibilités qu'il offre aux gestionnaires de données soucieux de produire des données de qualité et répondant aux principes FAIR.

Un environnement collaboratif pour faciliter la production de métadonnées

Pour réunir les informations contextuelles qui décrivent un jeu de données (i.e. les métadonnées), il est nécessaire d'interagir avec plusieurs acteurs d'une unité de recherche (responsable scientifique, collecteur, gestionnaire de données...). Chacun aura une vision précise d'un aspect du jeu de données que les autres n'ont pas. Ainsi, un point clé pour la saisie efficace de ces différentes informations consiste à offrir un outil mutualisé et facile d'accès qui permet l'édition collaborative des métadonnées. Cette saisie doit être initiée le plus en amont possible de la collecte des données, idéalement dès la rédaction du PGD, au moins pour ce qui concerne les métadonnées générales du projet dans laquelle s'inscrit la production de données.

Tableurs en ligne (Google Sheets)

La chaîne de traitement présentée se base sur l'outil Google Sheets pour la saisie des différents champs de métadonnées. Cette application offre plusieurs avantages :

- il est accessible à toute personne ayant un accès Internet, sans contrainte d'identifiants spécifiques et/ou de connexion à un réseau privé,
- il est également accessible hors ligne sur OS Android (85% des smartphones utilisés),
- c'est un outil couramment utilisé par un grand nombre de personnes, dans un cadre professionnel autant que privé,

⁵ <https://www6.inra.fr/datapartage/Documents-de-reference/Charte-Open-Access-de-l-INRA2>

⁶ http://cnig.gouv.fr/?page_id=8991

- la vitesse d'accès, la disponibilité et les capacités de stockage ne sont plus à démontrer,
- il propose une API interfaçable via différentes librairies R.

Néanmoins, il peut faire l'objet de réticences, voire de restrictions dans certains instituts. En effet, les serveurs hébergeant cet outil sont localisés hors du territoire français et il n'y a donc aucune garantie sur la protection et la réutilisation possibles des informations qui y sont stockées. Une réponse à ce questionnement légitime est que toutes les informations rassemblées dans une fiche de métadonnées - sauf exception particulière - sont *a priori* publiques, même si elles décrivent un jeu de données très confidentiel.

Il est tout à fait possible de remplacer cet outil collaboratif en ligne par un tableau CSV ou un autre format de données, partagé ou non. Des améliorations sont en cours de développement pour permettre le choix de la source de métadonnées (Google Sheets, CSV, base de données, autre outil de partage de fichiers).

R Studio Server

Toutes les fonctionnalités de traitements des éléments des métadonnées ont été codées en langage R. Ce langage présente de nombreux avantages :

- il est très largement utilisé dans la communauté scientifique nationale et internationale, principalement pour les traitements statistiques mais également pour le développement d'applications scientifiques, notamment celles liées à l'intelligence artificielle ;
- le logiciel R est libre et gratuit, ce qui garantit une réutilisation sans contrainte des codes produits ;
- la communauté de développeurs R est importante et très bien organisée ;

Une des limites pour une utilisation partagée du logiciel R est le nombre très important de librairies existantes, qui ne facilite pas la mise en place d'un paramétrage homogène du logiciel entre plusieurs partenaires. L'installation d'un serveur R Studio permet de passer outre cette limite, en mettant à disposition d'un collectif une interface pré-paramétrée du logiciel R, accessible via un simple navigateur Web.

Un serveur R permet également de déporter la charge de calcul généralement attribuée aux postes de travail personnels sur un serveur opérationnel en permanence.

Cet environnement collaboratif est le point clé du projet : tous les éléments nécessaires à la production des données sont accessibles à partir d'un navigateur Web.

Des librairies R pour générer des métadonnées normalisées

Un ensemble de librairies R ont été implémentées⁷ pour réaliser différentes actions liées aux métadonnées :

- *geometa* pour écrire / lire des métadonnées conformes aux standards OGC/ISO 19115 et 19139 (XML). Par défaut, la librairie vérifie la conformité des fiches écrites avec les normes ISO/OGC. Un module permettant de vérifier la conformité INSPIRE est également disponible ;
- *ows4R* pour utiliser les web services OGC à partir de R que ce soit pour l'interrogation des services de données (WFS) et métadonnées (CSW) mais également les opérations de transaction pour la gestion standardisée des (méta)données (insertion, mise à jour, suppression de métadonnées, CSW-T) ;
- *geonapi* pour interfacier le logiciel Geonetwork, et notamment publier les fiches de métadonnées générées avec *geometa* ;
- *geosapi* pour interfacier l'API du logiciel Geoserver, et notamment publier des flux WMS/WFS ;
- *zen4R* pour publier un jeu de données (accompagnées d'éléments de métadonnées principaux) dans l'entrepôt Zenodo depuis R en utilisant l'API de Zenodo.

⁷ <https://github.com/eblondel>

geometa

Le librairie geometa R fournit un modèle et des outils pour lire et écrire des éléments de (méta)données ISO / OGC pour les normes de métadonnées de base (ISO 19115, 19119). geometa prend également en charge la norme ISO 19110 (souvent partiellement ou non gérée par des interfaces graphiques) et son intégration dans les métadonnées de jeu de données ISO 19115. ISO 19110 permet aux utilisateurs de décrire la structure de données (ou dictionnaire de données), nécessaire pour une véritable réutilisation des jeux de données. La conformité ISO/OGC des métadonnées produites avec geometa peut être testée directement dans R. De récents développements financés par le projet R Consortium⁸ de la fondation Linux⁹ ont permis d'intégrer un validateur de métadonnée INSPIRE ainsi que la gestion de fiches de métadonnées multi-langues.

Projet: <https://github.com/eblondel/geometa>

CRAN: <https://cran.r-project.org/librairie=geometa>

ows4R

La librairie ows4R est quant à elle destinée à interagir avec les services web OGC dédiés aux données, avec la norme OGC Web Feature Service (WFS), et ceux dédiés aux métadonnées avec la norme OGC Catalogue Service for the Web (CSW). Il fournit des interfaces R équivalentes à celles de la librairie Python OWSLib ou ses équivalents Java Apache SIS/Geotoolkit et Geotools. Les flux de métadonnées (CSW) et les services Web de données vectorielles (WFS) sont pris en charge. L'intégration d'une interface pour les services de données raster (OGC Web Coverage Service – WCS) est en cours d'évaluation. Le service Web CSW-T permet la publication, la découverte, l'accès et la transaction (insertion, mise à jour, suppression) de fiches de métadonnées ISO 19139, y compris celles issues de geometa.

Projet: <https://github.com/eblondel/ows4R>

CRAN: <https://cran.r-project.org/librairie=ows4R>

geosapi

La librairie geosapi fournit une interface aux fonctionnalités du serveur géographique GeoServer, permettant ainsi de gérer un entrepôt de données géographiques, de configurer les couches géographiques, leur flux de données OGC associées (WMS, WFS), et les styles associés à ces couches géographiques. De cette manière, la plupart des tâches de gestion de données disponibles via l'interface graphique de GeoServer le sont également de manière programmatique avec R. Ceci permet d'automatiser la gestion/publication de jeux de données géographiques à partir de scripts R, et de synchroniser leurs mises à jour avec celles des métadonnées.

Projet: <https://github.com/eblondel/geosapi>

CRAN: <https://cran.r-project.org/librairie=geosapi>

geonapi

⁸ <https://www.r-consortium.org/>

⁹ <https://www.linuxfoundation.org/projects>

Projet : <https://github.com/eblondel/geonapi>
CRAN : <https://cran.r-project.org/librairie=geonapi>

Chacune des librairies décrites précédemment peut être utilisée de manière indépendante, selon les besoins. Néanmoins, il est très utile pour un gestionnaire de données de pouvoir combiner l'ensemble des librairies R dans une seule chaîne de traitement qui prend en charge les différentes opérations liées à la gestion des (méta)données.

- Les étapes de cette chaîne de traitement (présentées en Figure 1) répondent aux principaux objectifs de la FAIRisation des données.



Saisie des principaux éléments de métadonnées

En premier lieu, il s'agit de renseigner les éléments descriptifs d'un jeu de données. Les éléments retenus dans l'interface actuelle correspondent aux champs de la norme Dublin Core (ISO 15836). Ils correspondent aux champs minimaux requis par la norme géographique ISO 19115, permettant une identification rapide des ressources disponibles :

- identifiant,
- titre,
- résumé,
- contacts pour la ressource,
- mots-clés,
- date de création,
- type de ressource,
- format de la ressource,
- langue,
- liens,
- couverture spatiale,
- couverture temporelle,
- généalogie,
- droits associés.

L'approche proposée consiste à décrire chacun de ces éléments de métadonnées dans les colonnes d'un simple tableau (type CSV) qui intègre des conventions pour la syntaxe des éléments de métadonnées de cardinalité multiple (contacts, mots-clés...). L'édition peut se faire de différentes manières :

- en utilisant un tableur comme Libreoffice,
- directement en ligne avec Google Sheets,
- en ajoutant une table dédiée aux métadonnées dans le modèle d'une base de données.

L'exemple en Figure 2 présente une solution d'édition en ligne basée sur Google Sheets :

	A	B	C	D
1	Identifiant	Titre	Subject	Creator
2	Nom du fichier numérique des données sources (ou dossiers ou BDD) dans votre unité	Nom détaillé de la donnée ou du jeu de données ou de la collection des données ou de la série de données ou de l'ensemble des données; c'est ce nom qui apparaîtra sur l'interface pour visualiser vos métadonnées	Mots clés à aller chercher pour les mots clés "GENERAL" dans soit le thésaurus : - AGROVOC multilingues http://agrovoc.uniroma2.it/agrovoc/agrovoc/en?clang=fr : utiliser la barre de recherche en haut à droite de l'écran. Un petit menu déroulant à gauche de la barre de recherche vous permet d'indiquer la langue associée à votre recherche ET/OU - Agrovoc GEMET multilingue spécifique	Il s'agit ici de renseigner le producteur des données (owner), le publieur (publisher) et le gestionnaire des métadonnées (métada). Mettez une adresse mail à la place des adresses comme dans l'exemple, attention tout en gardant le même modèle: il faut impérativement que 3 noms apparaissent derrière chaque contribution (même si c'est à chaque fois la même adresse) = 3 adresse mail "owner", 2 adresses mail "publisher" et une adresse mail métada.
3	Fuseau_enquête agriculteurs	enquête audio agriculteurs pour le projet Fuseau	GENERAL=foncier, usages, protection des ressources, milieux aquatiques GEOGRAPHIC=France,	owner=christine.leger-bosch@irstea.fr;owner=nadia.guiffant@irstea.fr;publisher=christine.leger-bosch@irstea.fr;nadia.guiffant@irstea.fr;metada=nadia.guiffant@irstea.fr
4	DONNEES_CARTOGRAPHIE_LIGN EUX_COTEAUX	Cartographie fine des ligneux dans les bois des coteaux de Gascogne	GENERAL=arbre, données géoréférencées, forêt, distribution spatiale GEOGRAPHIC=France, Coteaux de Gascogne, Occitanie, Saint André	owner=alain.cabanettes@inra.fr;owner=wilfried.heintz@inra.fr;owner=wilfried.heintz@inra.fr;publisher=wilfried.heintz@inra.fr;publisher=alain.cabanettes@inra.fr;metadata=wilfried.heintz@inra.fr
5	STAGE_JULIE_SAUREL	Effets du climat et de la sylviculture sur la croissance en diamètre des chênes: cas de deux bois fragmentés des coteaux de Gascogne	GENERAL=Quercus petraea, Quercus rubra, Quercus pubescens, cerné, effet du climat, sylviculture, dendrochronologie.	owner=alain.cabanettes@inra.fr;owner=wilfried.heintz@inra.fr;owner=wilfried.heintz@inra.fr;publisher=wilfried.heintz@inra.fr;publisher=alain.cabanettes@inra.fr;owner=wilfried.heintz@inra.fr;publisher=wilfried.heintz@inra.fr;publisher=alain.cabanettes@inra.fr;metadata=laurent.bumel@inra.fr
6	THESE_OLA_MDAWAR	Les accrus de frêne (Fraxinus excelsior L.) à l'interface environnement/sylviculture dans les Pyrénées Centrales : distribution spatiale et temporelle	GENERAL=Fraxinus excelsior, colonisation, accroissement forestier, distribution spatiale, facteur biotique, facteur abiotique, sylviculture, cartographie	owner=alain.cabanettes@inra.fr;owner=wilfried.heintz@inra.fr;publisher=alain.cabanettes@inra.fr;metadata=laurent.bumel@inra.fr
7	promethee_nbr_feux_annee.	Nombre de feux à l'année , région sud	GENERAL= données attributaire GEOGRAPHIC=France, zone sud	owner=michel.mouliery@inra.fr;publisher=michel.mouliery@inra.fr;metadata=michel.mouliery@inra.fr

Figure 2 : Interface utilisateur de Google Sheets pour la saisie des métadonnées

Dans cet exemple, chaque ligne du tableur représente ainsi un jeu de données, chaque colonne correspondant à des éléments de métadonnées listés ci-dessus. Les éléments qui ont une cardinalité multiple font l'objet de

conventions d'écriture. Par exemple, des vignettes illustratives peuvent être ajoutées dans le champ relation en utilisant l'écriture `thumbnail@http://url_de_la_vignette`.

Paramétrage du workflow

L'ensemble des paramètres nécessaires à la bonne exécution de la chaîne de traitement est regroupé dans un fichier JSON (`workflow_configuration_Dublin_Core_gsheet.json`) qui doit être adapté à l'IDS qui va être alimentée. Ces paramètres permettent essentiellement de gérer les accès aux différents composants logiciels de l'IDS : Google Sheets, Geonetwork, Geoserver, Dataverse ...

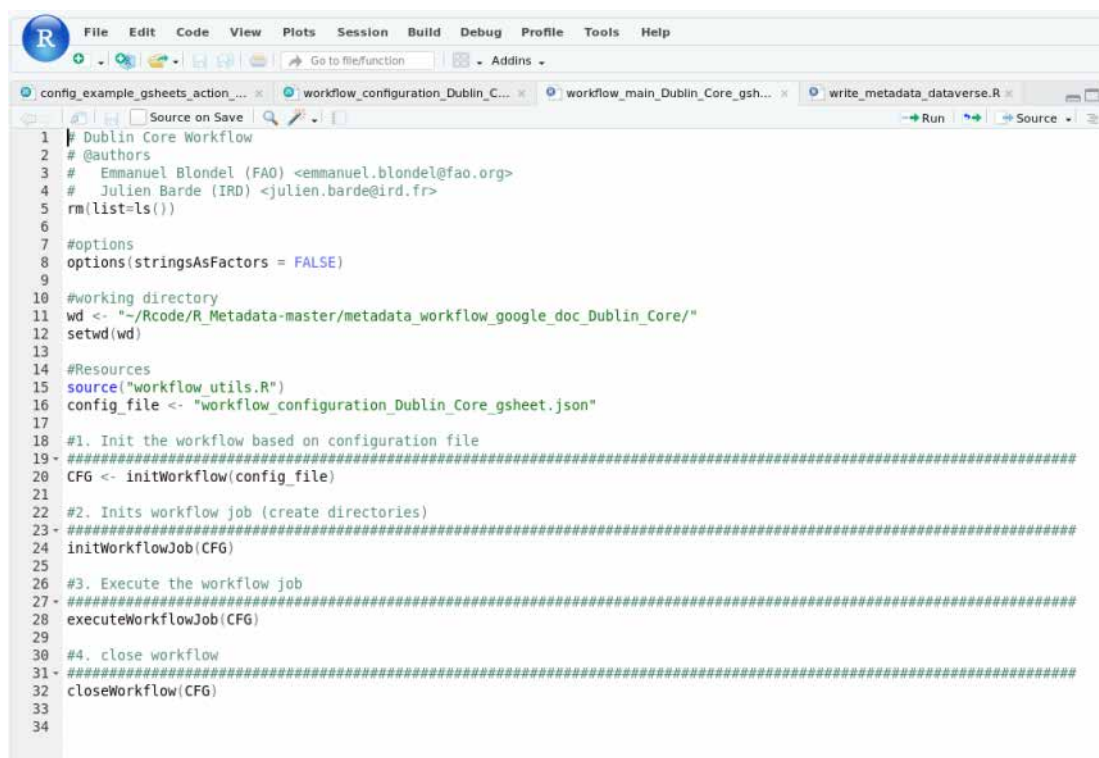
Exemple de paramètres :

```
"geonetwork": {  
  "url": "https://mygeonetwork.org/geonetwork",  
  "user": "*****",  
  "pwd": "*****",  
  "version": "3.0.4"}  
}
```

Exécution du workflow

La chaîne de traitement est lancée directement dans R en exécutant le fichier `workflow_main_Dublin_Core_gsheet.R`.

Nous présentons l'utilisation de la console RStudio dans la Figure 3 :



```
1 # Dublin Core Workflow  
2 # @authors  
3 # Emmanuel Blondel (FAO) <emmanuel.blondel@fao.org>  
4 # Julien Barde (IRD) <julien.barde@ird.fr>  
5 rm(list=ls())  
6  
7 #options  
8 options(stringsAsFactors = FALSE)  
9  
10 #working directory  
11 wd <- "~/Rcode/R_Metadata-master/metadata_workflow_google_doc_Dublin_Core/"  
12 setwd(wd)  
13  
14 #Resources  
15 source("workflow_utils.R")  
16 config_file <- "workflow_configuration_Dublin_Core_gsheet.json"  
17  
18 #1. Init the workflow based on configuration file  
19 #####  
20 CFG <- initWorkflow(config_file)  
21  
22 #2. Inits workflow job (create directories)  
23 #####  
24 initWorkflowJob(CFG)  
25  
26 #3. Execute the workflow job  
27 #####  
28 executeWorkflowJob(CFG)  
29  
30 #4. close workflow  
31 #####  
32 closeWorkflow(CFG)  
33  
34
```

Figure 3 : Console RStudio présentant le fichier principal du workflow

L'exécution du workflow lance plusieurs actions successives :

- le contenu des métadonnées est récupéré dans le tableur Google Sheets
- le XML de la métadonnée ISO 191139 est généré avec l'API geometa
- ce dernier est déposé dans notre instance Geonetwork via l'API geonapi ou bien ows4R (selon les paramètres activés)
- les flux WMS et WFS sont créés ou mis à jour dans geoserver via l'API geosapi (uniquement si les données sont stockées dans une table PostGIS)
- le jeu de données et la métadonnée associée sont déposés dans Dataverse via l'API dataverse. Un DOI est ainsi attribué au jeu de données.

Résultats

Lorsque toutes les étapes du workflow sont achevées, nous pouvons vérifier la publication de la ou des fiches de métadonnées dans le catalogue de métadonnées et/ou l'entrepôt de données désigné dans le fichier de configuration JSON. Si la fiche a bien été publiée, elle contient alors les différentes métadonnées (19115, 19119, 19110) et indique les protocoles d'accès disponibles pour exploiter les métadonnées et le jeu de données décrits de manière programmatique (CSW, WMS, WFS/WCS).

Conclusion et perspectives

Le workflow présenté ci-dessus est à ce jour en cours de consolidation pour une meilleure stabilité et durabilité (versions de l'outil, gestion suivie des correctifs et améliorations, et publication sous forme de librairie R) et ainsi favoriser une meilleure prise en main par les utilisateurs.

Ce travail se présente sous la forme d'un projet open-source et accessible sur <https://github.com/ebondel/geoflow>¹⁰. Parmi les améliorations en cours, nous pouvons citer la capacité d'étendre facilement la configuration du workflow pour une gestion personnalisée des différentes étapes de la chaîne de traitement.

Par exemple: utilisation d'un format d'entrée autre que Google Sheets comme source de métadonnées (e.g. CSV, base de données) ou contacts (e.g. CSV, base de données, annuaire LDAP), ou encore la facilité d'exécution du workflow par le biais d'une seule ligne de code R :

```
executeWorkflow("configuration.json").
```

La consolidation de ce workflow par les auteurs a pour perspective de fournir des outils clés en main pour l'utilisateur via un outil R d'orchestration et d'automatisation de tâches. Il permet la production de métadonnées minimales et/ou la publication de jeux de données associés sur des plateformes web diverses et complémentaires tels que les services web de (méta)données OGC (WMS, WFS, CSW), les entrepôts de données et d'attribution de DOIs (Dataverse, Zenodo) ou autres catalogues spécialistes (par exemple GBIF pour les données de biodiversité).

L'utilisateur avancé pourra personnaliser facilement l'outil (utilisation de sources/formats de données spécifiques, utilisation de scripts avancés).

¹⁰ <https://doi.org/10.5281/zenodo.3138920>

Une des perspectives sous-jacentes au travail de consolidation de ce workflow est la mise en place d'une interface graphique avec le framework Shiny R, pour la configuration et l'exécution du workflow, mais également pour l'édition du contenu des métadonnées.

Références bibliographiques

Blondel E (2018, August 22). geometa: Tools for Reading and Writing ISO/OGC Geographic Metadata in R (Version 0.3). Zenodo. <http://doi.org/10.5281/zenodo.1402245>

Blondel E (2018, August 14). geonapi: R Interface to GeoNetwork API (Version 0.2-0). Zenodo. <http://doi.org/10.5281/zenodo.1345102>

Blondel E (2018, September 12). geosapi: GeoServer REST API R Interface (Version 0.3). Zenodo. <http://doi.org/10.5281/zenodo.1414330>

Blondel E (2018, August 31). ows4R: R Interface to OGC Web-Services (Version 0.1-0). Zenodo. <http://doi.org/10.5281/zenodo.1407376>

Blondel E (2019, January 22). zen4R: R Interface to Zenodo REST API (Version 0.1). Zenodo. <http://doi.org/10.5281/zenodo.2547037>

Barde J (2018, September 13). juldebar/R_Metadata: v0.2 (Version v0.2). Zenodo. <http://doi.org/10.5281/zenodo.1415308>

Blondel E, Barde J, Heintz W (2019, April 3). geoflow: R engine to orchestrate and run geospatial (meta)data workflows (Version 0.1-beta). Zenodo. <https://doi.org/10.5281/zenodo.3138920>