Un workflow accessible pour une gestion simple et durable des données

Julien Barde, Emmanuel Blondel, Wilfried Heintz

Article dans le numéro spécial du cahier des techniques 2019

Gestion de données : rappel

Données

=

Gestion de données : rappel

Données sans métadonnées

=

Gestion de données : points de blocages

Simplifier la gestion des données !

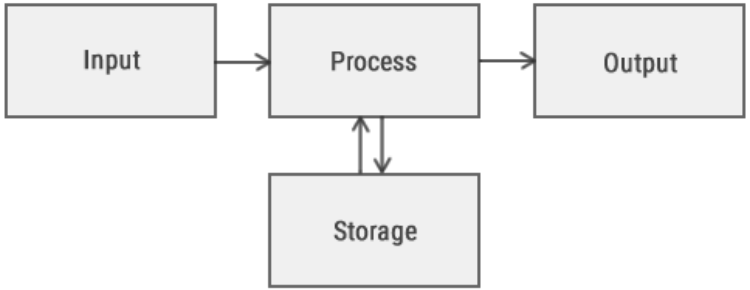
- **Favoriser** l'édition de métadonnées
  - Cibler en priorité les **métadonnées principales** (normes trop complexes)
  - ⇒ Dublin Core ou DataCITE
- **Shunter les interfaces graphiques** logicielles : Geonetwork, Geoserver… avec une prise en main qui fait fuir les utilisateurs
- **Automatiser** des éléments de métadonnées en lisant la donnée (si possible)

Dublin Core Metadata Initiative®
Making it easier to find information.

DataCite
FIND, ACCESS, AND REUSE DATA

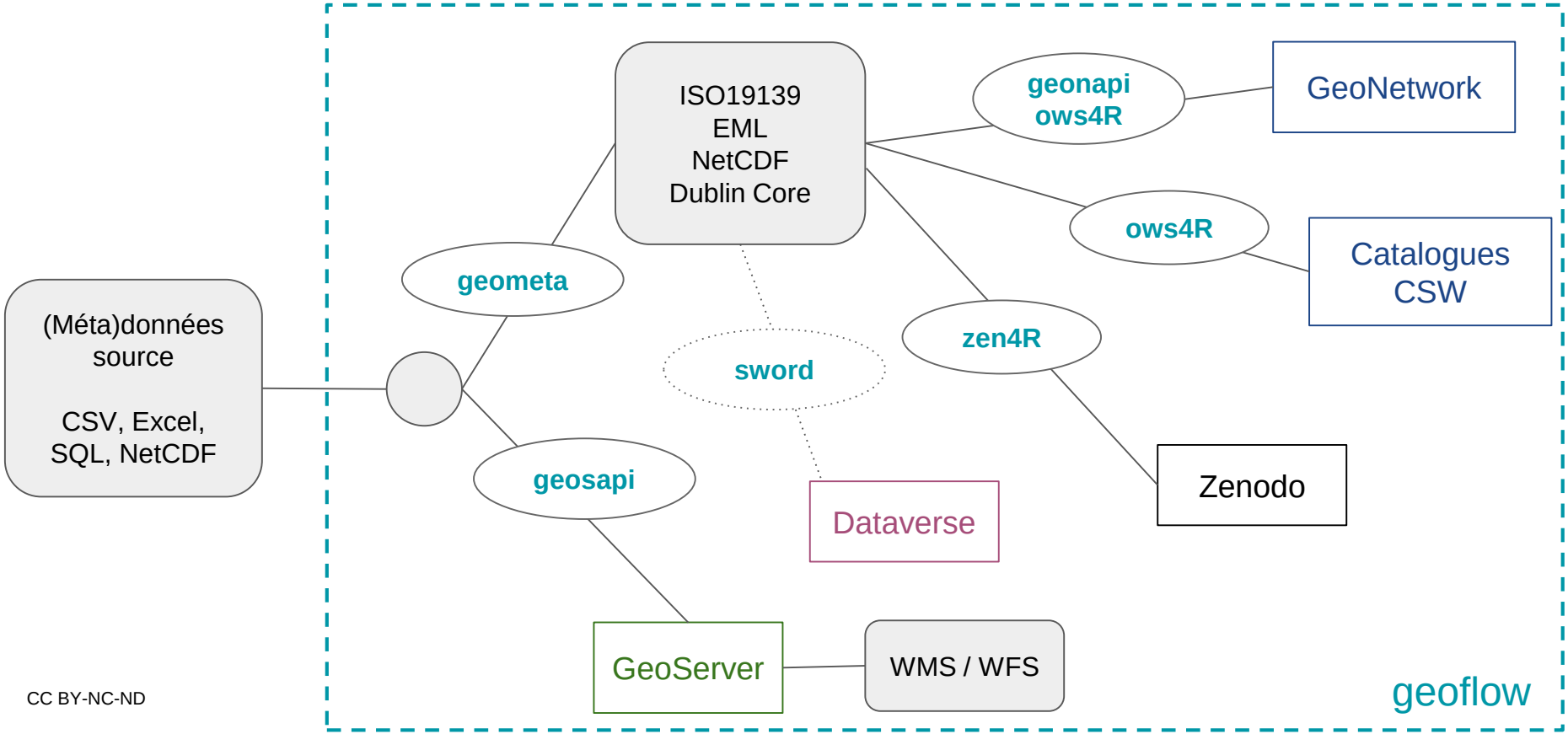
Workflow : objectifs de notre chaîne de traitements

Automatisation (ici en R) de tâches chronophages et récurrentes:



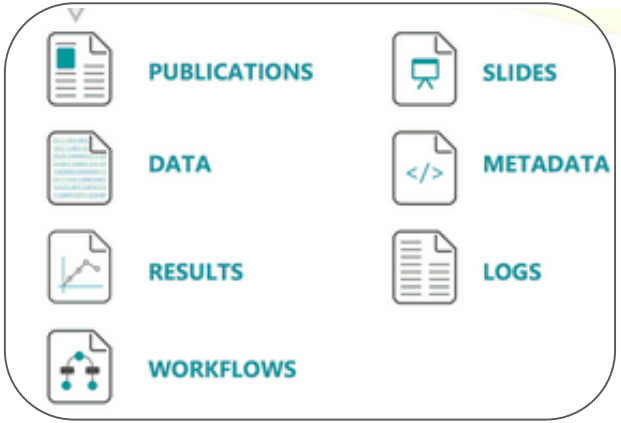
- Traitements, exemples d'**actions**:
  - Cataloguage de **métadonnées**
  - **Mapping** entre standards de métadonnées,
  - Publication de **(flux de) données**
  - Dépôts de (méta-)données avec attribution de **DOIs**
- **Flux** : du terrain aux entrepôts (Data Inra, Inspire, GBIF..)
- **Orchestration** des actions avec différents packages R

Workflow : vue d'ensemble



Workflow : exemples pour la donnée spatiale

- FAIRification des données basée sur les standards OGC
  - publication de **métadonnées** & **flux de données**
  - Attribution de **DOIs** et réinjection des DOIs dans les métadonnées
  - Lecture de métadonnées OGC depuis GN ou CSW
- **Visualisation** et **extraction** de données à partir des métadonnées
  - FAIR viewer (CSW+19115+19110+19119)
- **Publier** des research objects sur Zenodo



DOI 10.5281/zenodo.3138920

Geoflow : exécution

Un fichier de configuration config.json à adapter selon les besoins

- Deux fichiers essentiels:
  - Métadonnées: données décrites avec un modèle pivot (table de 15 colonnes DCMI avec des conventions de syntaxe). Stockées en CSV ou SQL, fichier, Google Spreadsheet et PostgreSQL (implicitement table “metadata”)
  - un fichier pour gérer les contacts (basé sur 19115, ~ EML)
- Lister les logiciels ou les infrastructures ciblés
  - Les connexions possibles:
    - Outils: Google Drive, PostgreSQL/PostGIS, Geonetwork, Geoserver (*vecteur*), Zenodo
    - Avec des protocoles standardisés (ex: OGC WFS et CSW)
  - En cours : Dataverse, Thredds
  - Perspectives : OGC WCS, GeoServer (*raster*), LDAP (source de contacts),

geoflow : FAIR workflow

Assurer les principaux services à partir de métadonnées riches

- Production de métadonnées conformes aux standards OGC data & services
  - 19115 (métadonnées)
  - 19110 (dictionnaire de données)
  - 19119 (services d’accès)
- Validation du contenu des métadonnées: un viewer FAIR (OpenFairViewer)
  - Métadonnées avec DOI
  - Accès aux données
  - Visualisation cartographique
  - Requêtage / Filtres sur les données sur des dimensions ou variables communes à différentes structures de données.

Conclusion

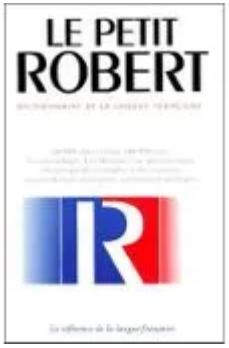
Beautifully Simple

- Méthode générique et indépendante basée sur les métadonnées et mise en oeuvre avec des outils simples
  - **Tableaux de saisie** simples (pour les utilisateurs)
    - **Métadonnées principales**: modèle pivot (DCMI) et mapping (19115, EML…)
    - **Structure de données:** basée sur 19110
    - **Contacts**
  - **Codes R** (pour les développeurs) avec automatisation si lecture possible de la donnée
- **Replicabilité** : gestion des configurations et logs, exécution en ligne…
- Workflow **customisable** pour gérer au cas par cas

Perspectives

- Interface **Shiny** pour l'exécution du workflow sans toucher ni R ni le JSON
- Renforcer l'**interopérabilité sémantique** : vocabulaires contrôlés
- Validation & Publication des métadonnées **INSPIRE**
- Gérer plus facilement les contacts avec des annuaires: LDAP & ORCID
- EML & Darwin Core vers OGC
  - Lire les archives sur un IPT
  - Mapping
  - Produire une version 19110 du Darwin Core
- Interagir avec des plateformes **cloud** (nextcloud, owncloud, EOSC)
- Alimenter d'autres documents "officiels" : PGD, Outils de suivi qualité ..

Workflow : recommandations



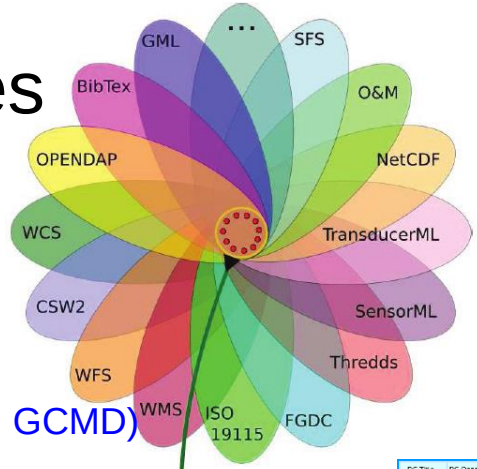
Se baser en amont sur des bonnes pratiques:

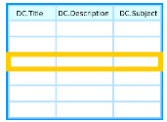
- Embarquer la **métadonnée** dans la donnée (si possible)
- Fournir un **dictionnaire de données** & un **catalogue de requêtes**
- **Interopérabilité sémantique** : vocabulaires contrôlés (+URI)
- **Réplicabilité** & **collaboration**: environnements mutualisés (serveurs / VRE) pour exécuter les workflows ou collaborer sur la saisie

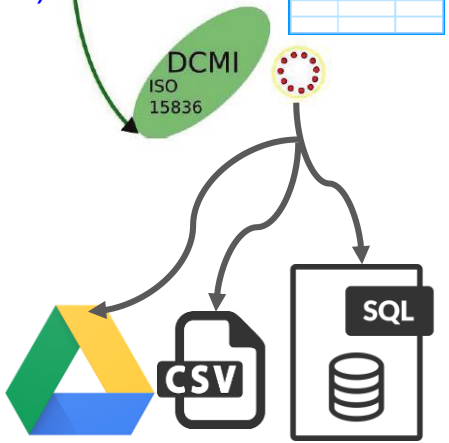
DOI 10.5281/zenodo.3138920

geoflow : modèle pivot de métadonnées

- **Identifier**: “string_identifier” + DOIs, URI, URNs..
- **Title**: “free text”
- **Description** : “prefix:free text” (abstract, purpose, additional info..)
- **Creator**: “role:person” / “role:email” + UID, ORCID / FOAF
- **Subject**: “thesaurus:keywords” or controlled vocabularies (eg GEMET, GCMD)
- **SpatialCoverage:** eWKT (dynamique si colonne “Data”)
- **TemporalCoverage**: ISO (dynamique si colonne “Data” avec dimension “time”)
- **Date**: controlled syntax
- **Type**: “free text”
- **Format**: “free text”
- **Language** : “code langue” (norme ISO 639: fre, eng..)
- **Relation** : “type:relation” or URLs, URIs...
- **Provenance**: “prefix:free text”
- **Rights**: “prefix:free text” or Creative Commons….
- **Data**: “prefix:URL” Rules to attach data



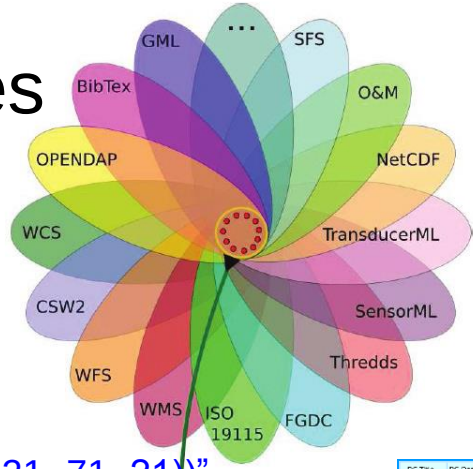


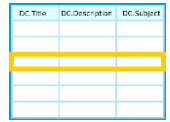


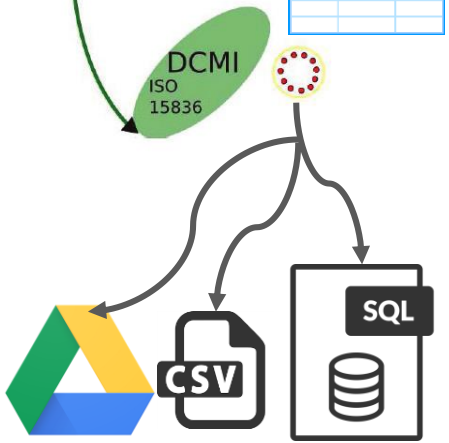
DOI 10.5281/zenodo.3138920

geoflow : modèle pivot de métadonnées

- **Identifier**: “my-geoflow-record0”
- **Title**: “My Title 0 - metadata only (no data associated)”
- **Description** : “abstract:My metadata description 0”
- **Creator**: “metadata:emmanuel.blondel1@gmail.com”
- **Subject**: “TH2:kwd1,kwd2,kwd3”
- **SpatialCoverage**: “SRID=4326;POLYGON((-71 -21,-71 28,14 28,14 -21,-71 -21))”
- **TemporalCoverage**: “2007/2010”
- **Date**: “2019-02-14”
- **Type**: “dataset”
- **Format**: DOIs
- **Language** : “eng”
- **Relation** : “http:website@http://somelink/website”
- **Provenance**: “process:rationale1[description1],rationale2[description2],..”
- **Rights**: “useConstraint:copyright”
- **Data**: “source:cwp_report.pdf@http://www.fao.org/3/i7805en/I7805EN.pdf”







liste Fablab INRA

https://groupes.renater.fr/sympa/arc/fablab