geoflow : workflow R pour gérer les données spatiales

DOI 10.5281/zenodo.3138920

Julien Barde, Emmanuel Blondel, Wilfried Heintz

https://tinyurl.com/y22zh9w5

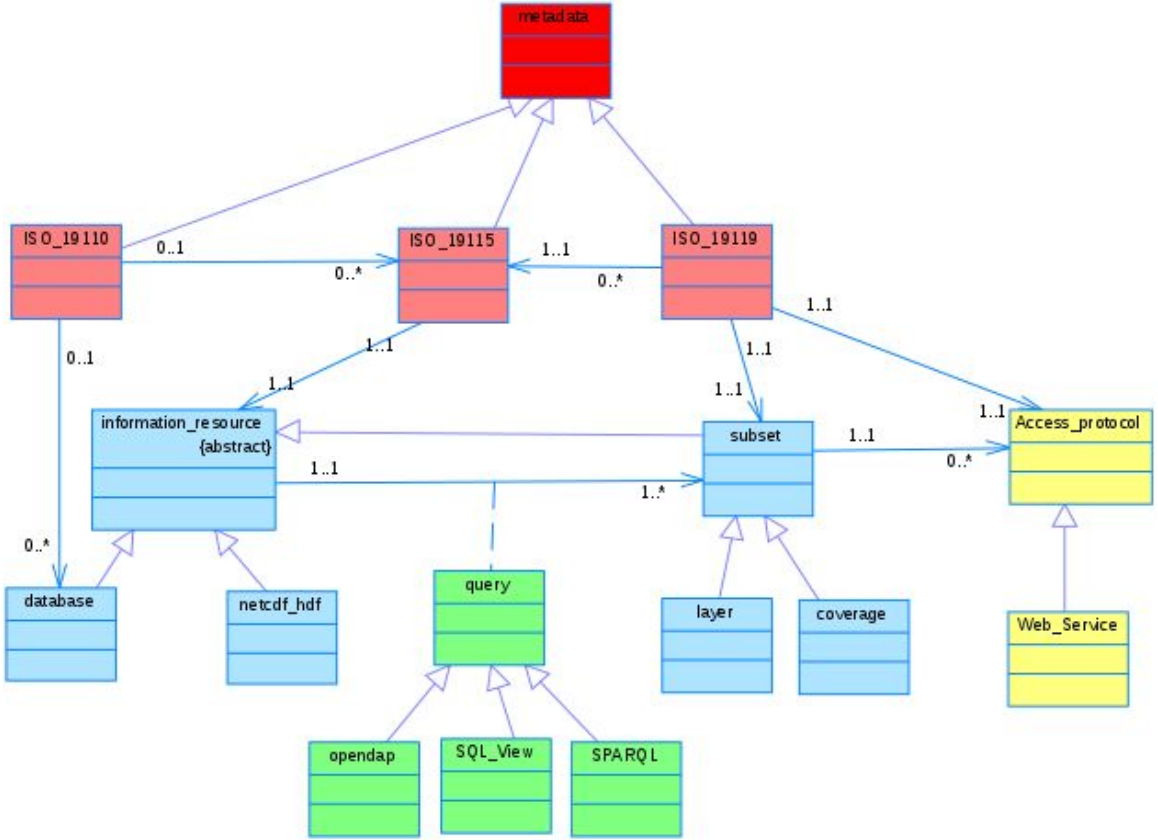
Plan de la présentation

- (Infrastructure de) Données Spatiales (SDI) et points de blocages
- Workflow pour la donnée spatiale
  - Objectifs
  - Exemples
  - Vue d’ensemble
- Matériel et méthodes
- Geoflow:
  - bonnes pratiques
  - matériel et méthodes
  - modèle pivot de métadonnées
  - exécution et fichier de configuration config.json
  - mapping de métadonnées
  - passion tuning

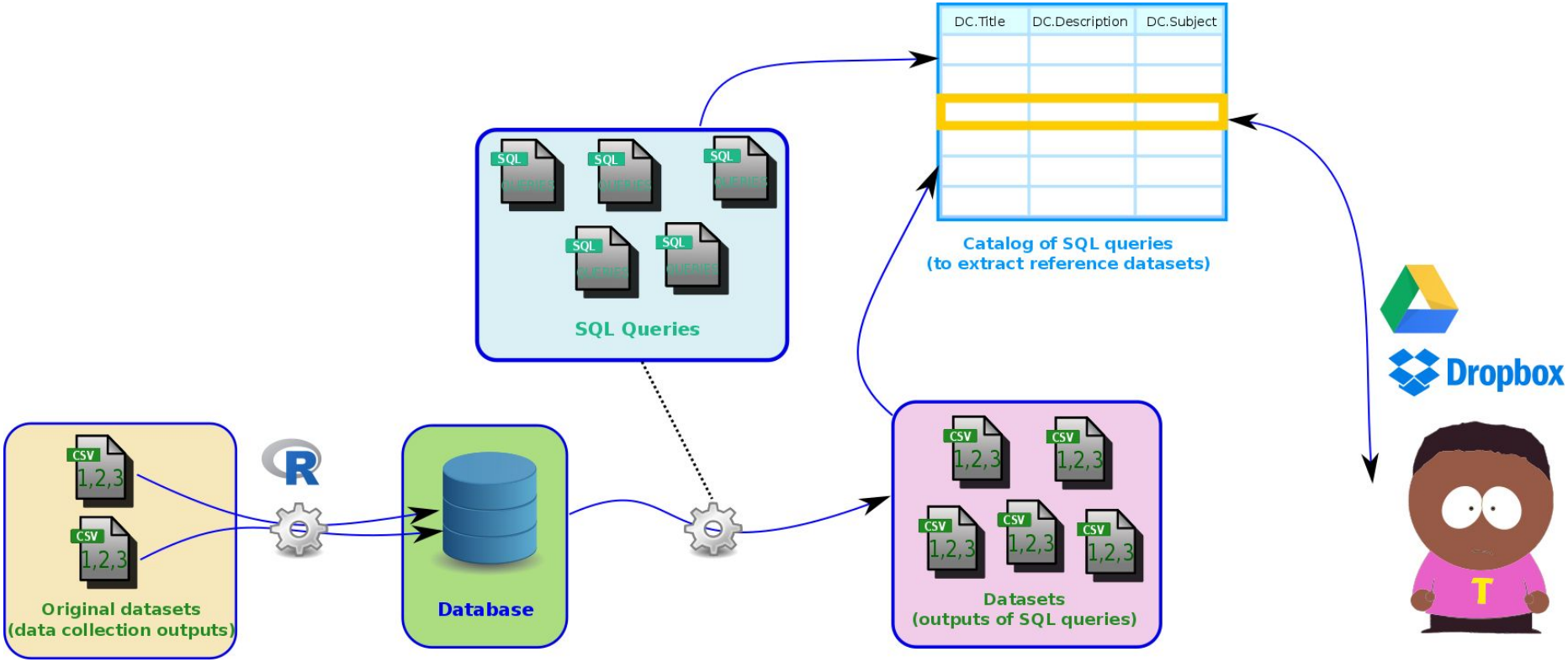
Historique

- Importance de R dans notre communauté mais rien pour les SDI
- H2020 BlueBridge
  - VRE
  - Datawarehouse postgis
  - Émergence du FAIR
  - => automatisation de services FAIR sur une base de données en R
- Packages R (Emmanuel Blondel):
  - geometa, geonapi, geosapi, ows4R, zen4R, geoflow
  - Workshop 2018 => améliorer la qualité des codes (R Metadata)
  - Financement R consortium
- Nouveaux projets et cas d’utilisation: Blue Cloud..

Données Spatiales: sources et jeux de données

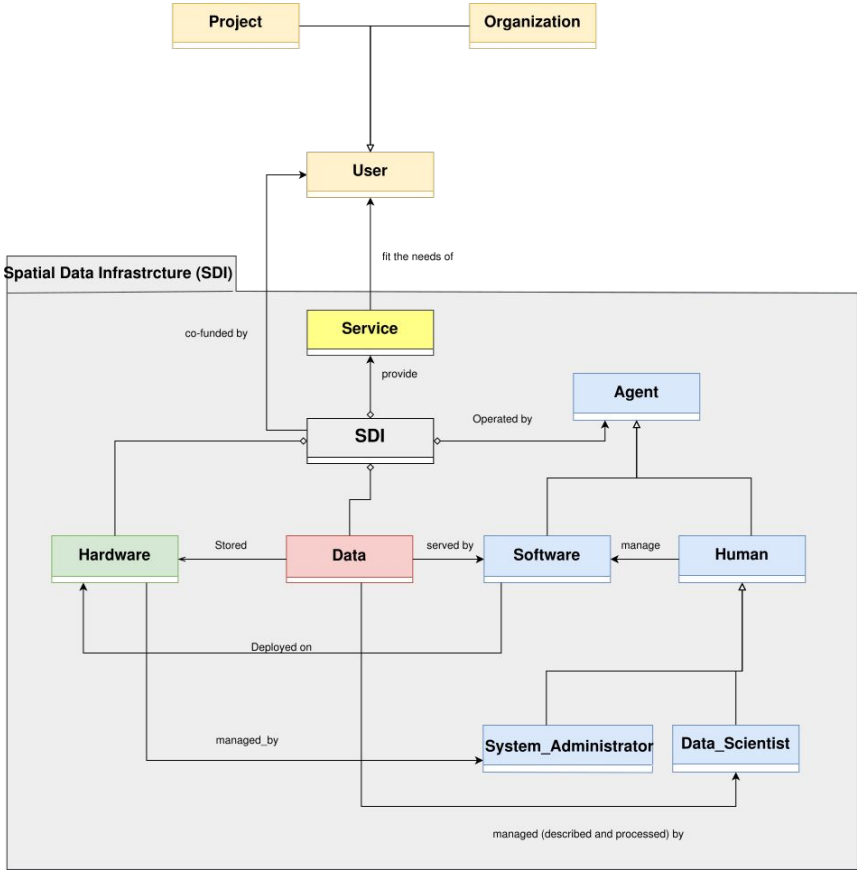


Données Spatiales: sources et jeux de données



Données Spatiales: Infrastructures (SDI)

- Sources de (méta-)données:
  - Fichiers: CSV, shapefiles..
  - SQL / SFS,
  - NetCDF & OPeNDAP.
- Process / Actions / Services:
  - Serveurs de (méta-)données:
    - Geonetwork, pycsw
    - Postgres & Postgis,
    - Geoserver (à venir: Thredds..),
  - Serveurs de traitements: RStudio, Jupyter
- Lien avec les infrastructures extérieures:
    - Zenodo, GBIF, Pangaea
    - INSPIRE,
    - D4science, EOSC,



Données Spatiales: points de blocages

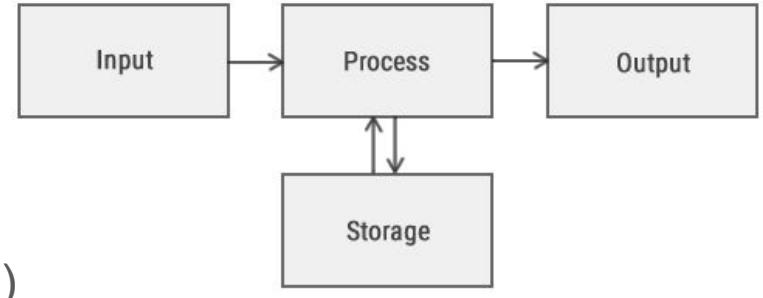
Simplifier la gestion des données:

- Absence / perte de métadonnées,
- Ecriture / Saisie de métadonnées:
  - Création et mises à jour de métadonnées (en impliquant les producteurs et les développeurs),
  - Mapping de métadonnées: publication “simultanée” dans différents catalogues pour éviter les saisies multiples: DCMI, OGC, EML, datacite (SWORD API pour Dataverse, Zenodo API..),
- Cibler en priorité les métadonnées essentielles (normes trop complexes),
- Shunter les interfaces graphiques complexes: Geonetwork, Geoserver…avec une prise en main qui fait fuir les utilisateurs,
- Automatiser des éléments de métadonnées en lisant la donnée (si disponible),
- Validation (ex: INSPIRE)

Workflow: objectifs d’une chaîne de traitements

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

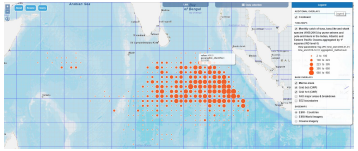
- Traitements, exemples d’**actions**:
  - Création & Publication de métadonnées (OGC..),
  - Mapping entre standards de métadonnées,
  - Publication de (flux de) données (eg WMS, WFS..),
  - Dépôts de (méta-)données avec attribution de DOIs
- Du labo aux infrastructures..(INSPIRE, GBIF..)
- Orchestration des traitements (“**actions**”) avec différents packages R:
  - OGC: geometa, ows4R et associés (geonapi, geosapi, geoserver, RpostgreSQL, ncdf4..)
  - Zenodo (zen4R), GBIF (eml, rgbif), Dataverse (dataverse)
  - En discussion: rorcid, LDAP, gsheet / googledrive, owncloud (ie B2DROP), redland / rrdf, exifr
- Execution en ligne (dans un VRE / EOSC) avec RStudio, Jupyter Notebook



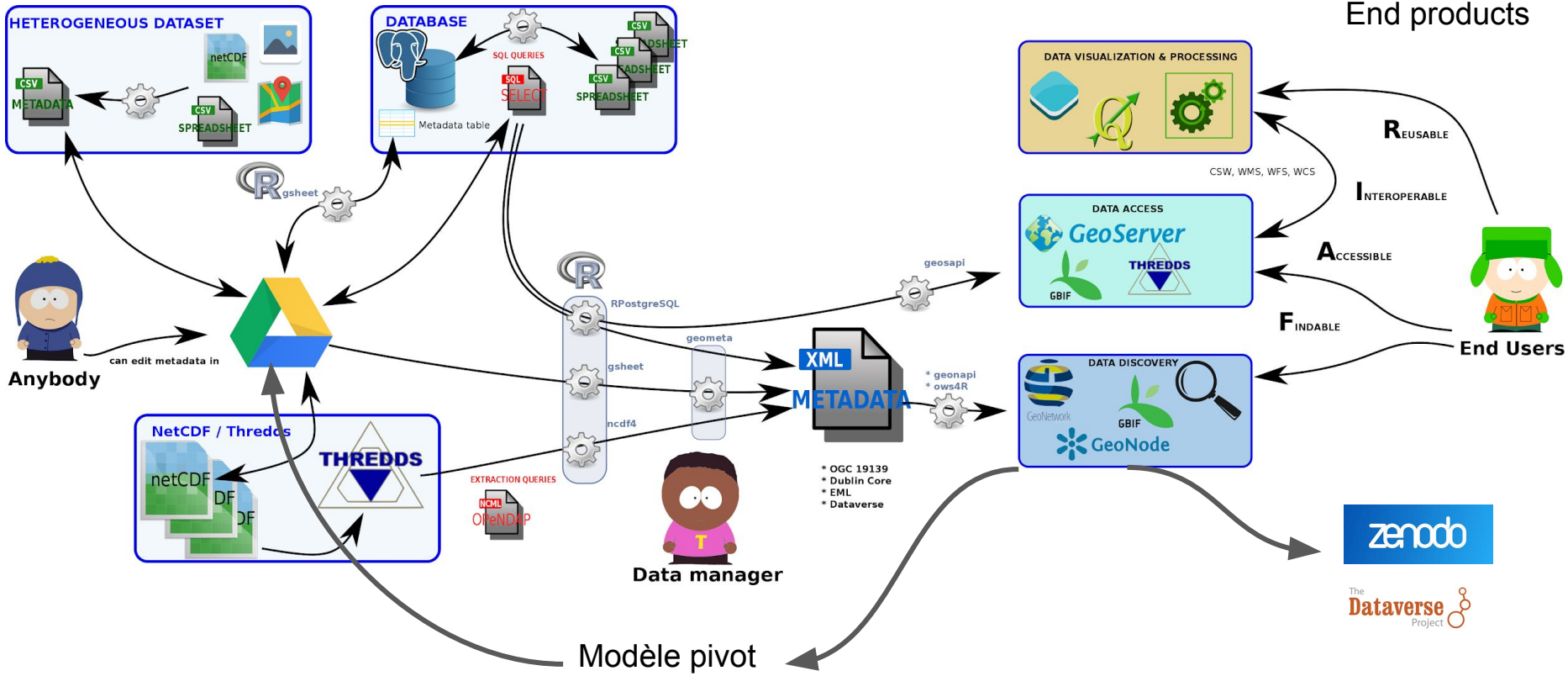
Workflow: exemples pour la donnée spatiale

- “FAIRification” des données basée sur les standards OGC:
  - publication de métadonnées & flux de données (dans SDI)
    - Définition d’un modèle pivot tabulaire (basé sur DCMI) vers OGC, Datacite, EML..knitr/Rmd
    - Modèle pivot implémenté en CSV, Google sheet, SQL / Postgis,
    - Traitements de lots de photos (drone, terrain..): modèle pivot et exif.
  - Attribution de DOIs et réinjection des DOIs dans les autres métadonnées
  - Lire des métadonnées OGC depuis GN ou CSW ou catalogues Thredds (NetCDF-CF ou NCML) pour charger le modèle pivot..
- Visualisation et extraction de données à partir des métadonnées: FAIR viewer (CSW+19115+19110+19119)
- Perspectives, exemples:
  - Publication de lots de données anciennes sur Zenodo,
  - Lire le contenu de l’IPT / GBIF et publier la métadonnée EML dans GN, la donnée dans GS,
  - Descriptions de lots de photos (missions terrain / drones)

Workflow: vue d'ensemble



End products



geoflow

Se baser en amont sur des bonnes pratiques:

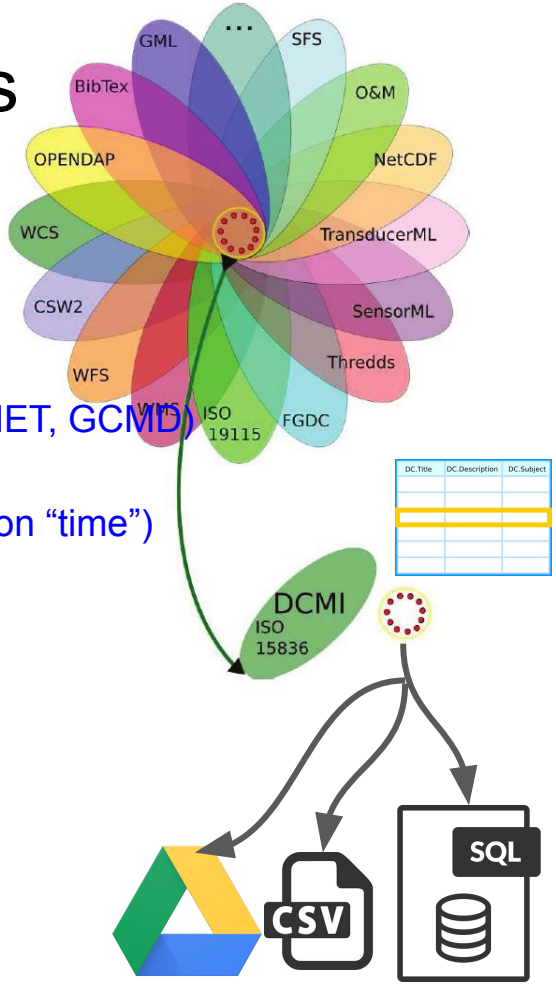
- Métadonnée embarquée dans la donnée (si possible => producteurs..)
- Fournir un dictionnaire de données et un catalogue de requêtes
  - Adopter des structures de données standardisées pour les cas classiques: Darwin Core / Measurement and facts: occurrences d'espèces et biométrie, O&M, OBOE, Conventions-CF
  - Exemple du NetCDF (ou NCML pour gérer la métadonnée à part ou des fichiers virtuels)
    - conformité avec les conventions CF pour les métadonnées et la structure des données (ce que demande thredds, gdal et donc qgis..),
    - serveur OPeNDAP (Thredds) => multiples protocoles d'accès (si conformité CF),
- Interopérabilité sémantique: URI & vocabulaires contrôlés,
- Réplicabilité & collaboration: environnements (serveurs / VRE) pour exécuter les workflows ou collaborer sur la saisie.

geoflow: matériel et méthodes

- Modèle pivot de métadonnées basé sur Dublin Core
- Approche mixte en fonction des profils utilisateurs:
  - producteurs de données: GUIs minimalistes et collaboratives pour que les métadonnées soient éditées facilement,
  - gestionnaires de données: exécution des traitements (actions) en R.
- Viser les sources de données classiques:
  - Shapefiles,
  - SQL/SFS: ajouter une table “metadata” au modèle,
  - NetCDF (codes à intégrer): conventions CF dans le NetCDF ou en dehors (NCML)
- Privilégier les applications standards:
  - CSW-T en priorité ou APIs natives de geonetwork (rq: problème avec geonode), WFS-T..
- Mise en oeuvre en R: le package geoflow orchestre des actions en coordonnant les actions déléguées à différents packages (dépendances),
- Travailler en ligne sans installation “Desktop”: VREs de Blue Cloud (EOSC)

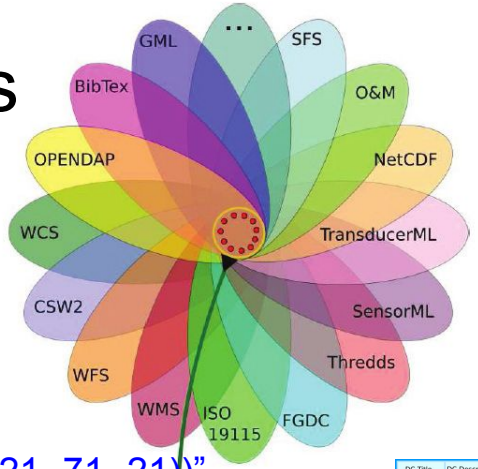
geoflow: modèle pivot de métadonnées

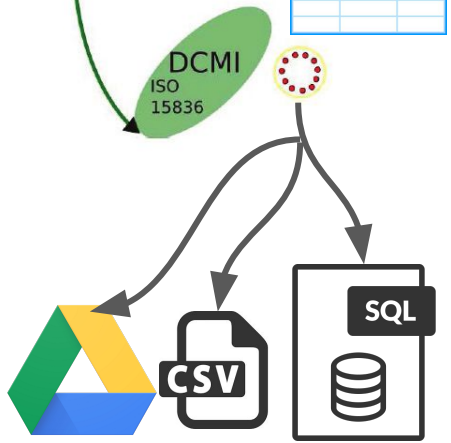
- **Identifier**: “string_identifier” + DOIs, URI, URNs..
- **Title**: “free text”
- **Description** : “prefix:free text” (abstract, purpose, additional info..)
- **Creator**: “role:person” / “role:email” + UID, ORCID / FOAF
- **Subject**: thesaurus:keywords KVP or controlled vocabularies (eg GEMET, 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 eg fre, eng → norme ISO 639
- **Relation** : type:relation KVP or URLs, URIs...
- **Provenance**: “prefix:free text”
- **Rights**: “prefix:free text” or Creative Commons….
- **Data**: “prefix:free text” Rules to attach data



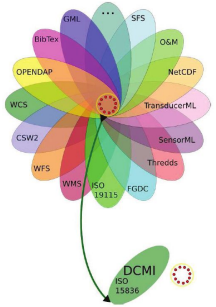
geoflow: modèle pivot de métadonnées

- **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”





geoflow: modèle pivot de métadonnées (CSV..)



DC.Title	DC.Description	DC.Subject

geoflow_entities

File Edit View Insert Format Data Tools Add-ons Help Last edit was made yesterday at 8:19 PM by Emmanuel Blondel

Share

Identifier

Identifier	Title	Description	Subject	Creator	Date	Type	Language	SpatialCoverage
my-geoflow-record0	My Title 0 - metadata onl	abstract:My metadata descripti purpose:Some purpose for this info:some supplemental info 0	TH1@http://somevocab?thesaurus=th1:kwd1@http://somevocab?keyword=kw TH2:kwd1,kwd2,kwd3_ TH3:kwd1,keyword 2@http://someref/kwd2,keyword 3	metadata:emmanuel.blondel1@gmail.com_ owner:emmanuel.blondel1@gmail.com_ pointOfContact:julien.barde@ird.fr,wilfried.heintz@inra.fr	2019-02-14	dataset	eng	SRID=4326;POLYGON(
my-geoflow-record1	My Title 1 - with a shapef	abstract:My metadata descripti purpose:Some purpose for this info:some supplemental info 1	TH1@http://somevocab?thesaurus=th1:kwd1@http://somevocab?keyword=kw TH2:kwd1,kwd2,kwd3_ TH3:kwd1,keyword 2@http://someref/kwd2,keyword 3	metadata:emmanuel.blondel1@gmail.com_ owner:emmanuel.blondel1@gmail.com_ pointOfContact:julien.barde@ird.fr,wilfried.heintz@inra.fr	2019-02-14	dataset	eng	SRID=4326;POLYGON(
my-geoflow-record2	My Title 2 - with a shapef	abstract:My metadata descripti purpose:Some purpose for this info:some supplemental info 2	TH1@http://somevocab?thesaurus=th1:kwd1@http://somevocab?keyword=kw TH2:kwd1,kwd2,kwd3_ TH3:kwd1,keyword 2@http://someref/kwd2,keyword 3	metadata:emmanuel.blondel1@gmail.com_ owner:emmanuel.blondel1@gmail.com_ pointOfContact:julien.barde@ird.fr,wilfried.heintz@inra.fr	2019-02-14	dataset	eng	SRID=4326;POLYGON(
my-geoflow-record3	My Title 3	abstract:My metadata descripti purpose:Some purpose for this info:some supplemental info 3	TH1@http://somevocab?thesaurus=th1:kwd1@http://somevocab?keyword=kw TH2:kwd1,kwd2,kwd3_ TH3:kwd1,keyword 2@http://someref/kwd2,keyword 3	metadata:emmanuel.blondel1@gmail.com_ owner:emmanuel.blondel1@gmail.com_ pointOfContact:julien.barde@ird.fr,wilfried.heintz@inra.fr	2019-03-07	dataset	eng	
my-geoflow-record4	My title 4	abstract:My metadata descripti purpose:Some purpose for this info:some supplemental info 4	TH1@http://somevocab?thesaurus=th1:kwd1@http://somevocab?keyword=kw TH2:kwd1,kwd2,kwd3_ TH3:kwd1,keyword 2@http://someref/kwd2,keyword 3	metadata:emmanuel.blondel1@gmail.com_ owner:emmanuel.blondel1@gmail.com_ pointOfContact:julien.barde@ird.fr,wilfried.heintz@inra.fr	2019-02-14	dataset	eng	SRID=4326;POLYGON
my-geoflow-record5	My title 5	abstract:My metadata descripti purpose:Some purpose for this info:some supplemental info 5	TH1@http://somevocab?thesaurus=th1:kwd1@http://somevocab?keyword=kw TH2:kwd1,kwd2,kwd3_ TH3:kwd1,keyword 2@http://someref/kwd2,keyword 3	metadata:emmanuel.blondel1@gmail.com_ owner:emmanuel.blondel1@gmail.com_ pointOfContact:julien.barde@ird.fr,wilfried.heintz@inra.fr	2019-02-14	dataset	eng	

example | definition | Sheet3

Explore

geoflow, exécution: executeWorkflow("config.json")

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 Postgres (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, Postgres/Postgis, Geonetwork, Geoserver (*vecteur*), Zenodo
    - Avec des protocoles standardisés (ex: OGC WFS et CSW)
  - “A migrer”: Dataverse, Thredds, autres??
  - “Perspectives”: OGC WCS, GeoServer (*raster*), LDAP (source de contacts),
  - Problème avec Geonode (CSW-T)

geoflow: fichier de configuration [config.json]

```
{
  "id": "my-workflow",
  "mode": "entity",
  "metadata": { <metadata sources defined here> },
  "software": [ <pieces of software defined here> ],
  "actions": [ <actions defined here>  ],
  "profile": { <global profile (metadata) defined here> },
  "options": { <global options defined here> },
}
```

geoflow: fichier de configuration config.json

```
"metadata": {
  "entities": {
    "handler": "gsheet",
    "source": "https://docs.google.com/spreadsheets/d/1iG7i3CE0W9zVM3QxWfCjoYbqj1dQvKsMnER6kqwDiqM/edit?usp=sharing"
  },
  "contacts" : {
    "handler": "gsheet",
    "source": "https://docs.google.com/spreadsheets/d/144NmGsikdIRE578IN0McK9uZEUHZdBuZcGy1pJS6nAg/edit?usp=sharing"
  }
}
```

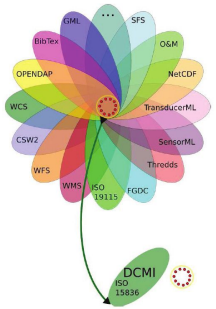
geoflow: fichier de configuration [config.json]

```
"metadata": {
  "entities": {
    "handler": "csv",
    "source": "~/metadata.csv”
  },
  "contacts" : {
    "handler": "csv",
    "source": "~/contacts.csv”
  }
}
```

geoflow: fichier de configuration [config.json]

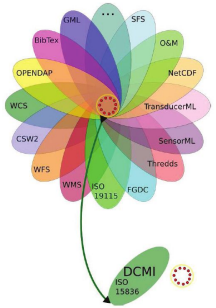
```
"metadata": {
  "entities": {
    "handler": "dbi",
    "source": "metadata”
  },
  "contacts" : {
    "handler": "csv",
    "source": "~/contacts.csv”
  }
}
```

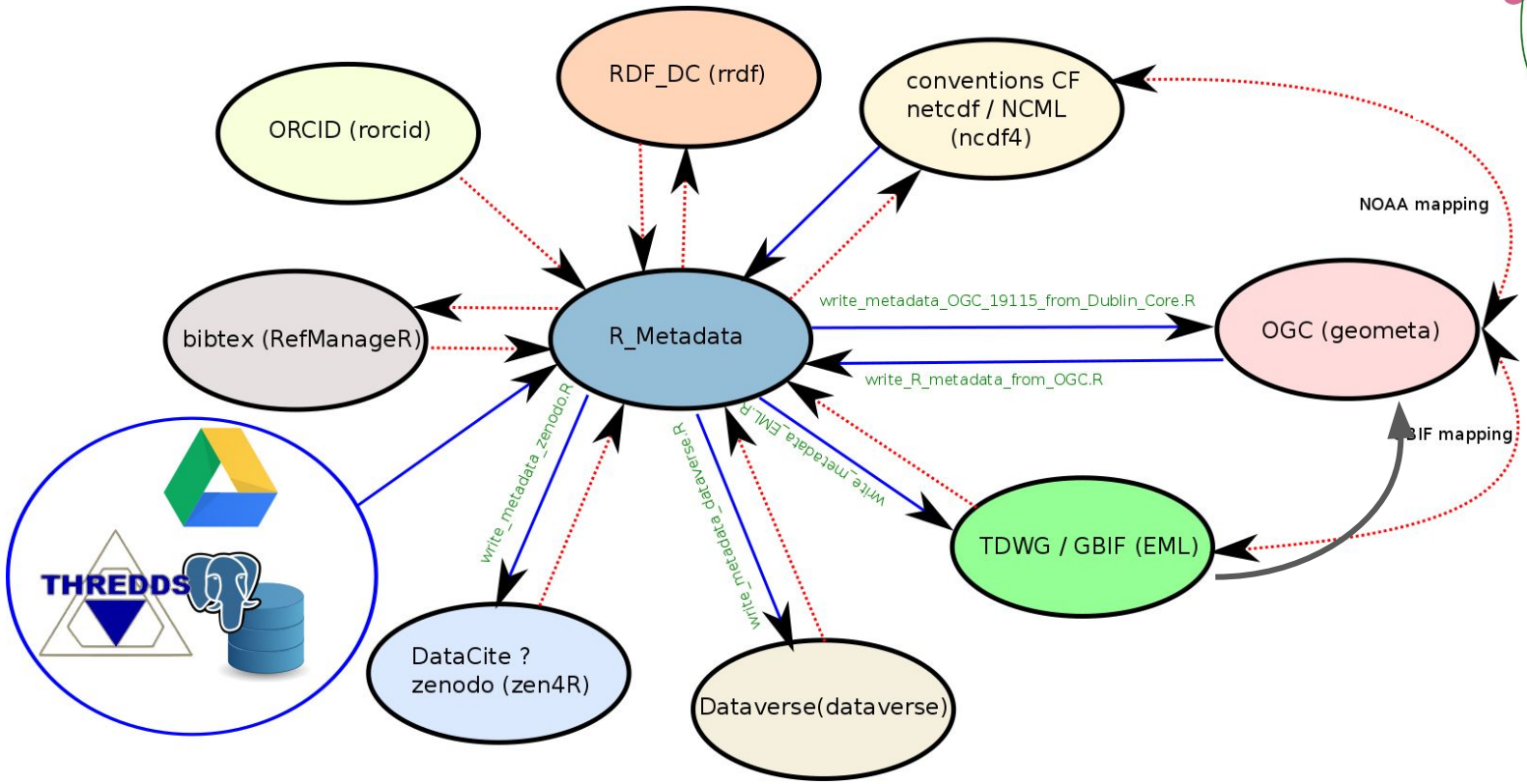
geoflow: mapping de métadonnées



- Les normes ciblées:
  - DCMI, Datacite (Zenodo API), ORCID, sword (API Dataverse)
  - OGC: Formats (19115/19110/19119) et Protocoles (CSW/WMS/WFS/WCS/OPeNDAP)
  - EML & Darwin Core (structure de données 19110 par défaut)
- Deux approches sont expérimentées en parallèle:
  - Directement d’une norme à une autre norme (via geometa):
    - 19115 (geometa) vers EML, …
  - Du modèle pivot vers les autres normes (codes R_metadata à migrer dans geoflow):
    - Avantages: export vers toutes les normes gérées à partir du modèle pivot
  - Inconvénients: restreint aux éléments gérés par le mapping avec le modèle pivot

geoflow: mapping de métadonnées (packages)





geoflow (spreadsheet, GN, GS) => fiche simple

Catalogue de métadonnées de l'UMR MARBEC / OSU-OREME | Rechercher | Visualiser | S'identifier | Français

Retour à la recherche

Carte de sensibilité du complexe récifal de Foulpointe (Madagascar)

Mis à jour : il y a 2 ans

La carte de sensibilité écologique récifale de Foulpointe à Madagascar. Elle a été réalisée dans le cadre des programmes environnementaux de la Commission de l'Océan Indien portant sur l'étude cartographique et sédimentologie du récif de Foulpointe.

A Madagascar, comme dans les autres pays de la COI, la politique environnementale revêt un aspect prioritaire par le lien direct qui existe entre la qualité de l'environnement marin et les activités socio-économiques intéressant la bande côtière tourisme , pêche, urbanisation, etc.

L'étude dont est issue cette carte entre dans cette politique. Son objectif principale est de comprendre les causes du phénomène d'hypersédimentation du récif de Foulpointe en vue de sa gestion durable. Elle a été réalisé pour le compte du Ministère de l'Environnement, des Eaux et Forêt par la Division de la Recherche et d'Application de la Gestion Intégrée des Zones Côtières du Centre National de Recherche sur l'Environnement (CNRE) avec l'appui technique de l'Agence pour la Recherche et la Valorisation Marine (ARVAM). Le financement a été assuré par la COI via les reliquats du PRE/COI-UE.

La présente carte classifie les différentes entités du récif selon leur sensibilité écologique qui prend en considération de nombreux paramètres faisant appel à des caractéristiques structurelles et/ou fonctionnelles des écosystèmes.

Under development

Téléchargements et liens

Lien de la référence zotero du document associé

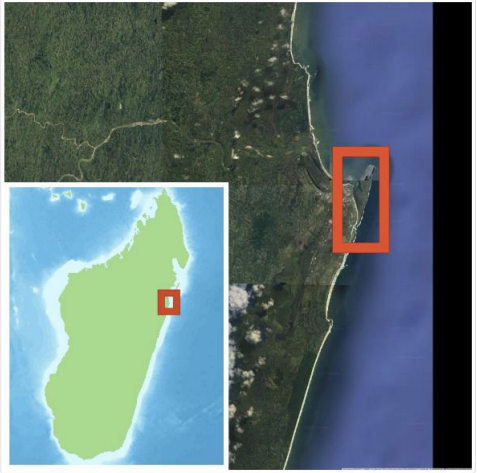
Cartographie et bilan diagnostic du complexe récifal Foulpointe. (Cote Est de Madagascar)

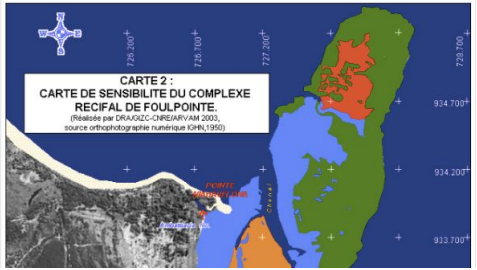
Ouvrir le lien

À propos de cette ressource

Catégories	
Mots-clés	• Récifs coralliens • Récifs • Sensibilité • Ecologie • Océan Indien • Madagascar • Foulpointe • Commission de l'Océan Indien • Programme Régional Environnement PRE/COI-UE • CNRE • ARVAM • Biota • Oceans • Environment
Langue	• Anglais

Aperçu





geoflow (Postgis, GN, GS) => Tuna Fisheries

zenodo

zenodo | Search | Upload | Communities | Log in | Sign up

February 1, 2018

Dataset | Open Access

Global monthly catch of tuna, tuna-like and shark species (1950-2015) aggregated by 1° or 5° squares (IRD level 2)

Paul Taconet; Emmanuel Chassot; Julien Barde

Data collector(s)

Fabio Fiorellato; Carlos Palma; Peter Williams; Nickolas Vogel; Colin Millar

This dataset lists global catch of tuna, tuna-like and shark species from 1950 to 2015. Catches are stratified by month, species, gear, vessel flag reporting country, fishing mode (i.e. type of school used), area (mainly 1° or 5° square) and unit of catch (weight or number). This dataset was computed using public domain catch-and-effort datasets released by the five tuna Regional Fisheries Management Organizations: the Indian Ocean Tuna Commission (IOTC), the International Commission for the Conservation of Atlantic Tunas (ICCAT), the Inter-American Tropical Tuna Commission (IATTC), The Western and Central Pacific Fisheries Commission (WCPFC), the Commission for the Conservation of Southern Bluefin Tuna (CCSBT).

IRD Level 2 stands for the processes applied to the primary datasets by the French National Research Institute for Sustainable Development (IRD) to generate the dataset. These processes for level 2 at global scale are :

- Catch-and-effort data are disseminated in such way that redundancy may exist between the various datasets released, or that dimensions may be split over the datasets for some strata. To cope with these issues and get one single and more complete possible value of catch per stratum (i.e. with all the available dimensions), these datasets had to be merged in specific ways - i.e. not simply merging them but removing the duplicated strata or reassembling the strata with all the available dimensions split over the datasets.
- Values expressed in weight were kept and numbers were converted into weight using simple conversion matrices (A. Fonteneau, pers. com). These conversion factors depend on the species, the gear, the year and the main geographical area (equatorial or tropical). They were computed from the Japanese and Taiwanese size-frequency data as well as from the Japanese total catches and catch-and-effort data. Some data might not be converted at all because no conversion factor exists for the stratum: those data were kept and the unit of catch was set to Number of fishes harvested. Some data might not be converted at all because no conversion factor exists for the stratum: those data were kept and the unit of catch was set to Number of fishes harvested.
- Geo-referenced catches were raised to the total catches.
- Data located at land or without any spatial information were equally redistributed on data at sea on areas with same characteristics (same year, month, gear, flag, species, type of school).
- In the overlapping zone between the IATTC and the WCPFC, only data from the IATTC was kept (i.e. for this specific zone, data from the WCPFC was removed). IATTC data was kept rather than WCPFC's because the information on vessel flag reporting country is available in IATTC datasets contrary to WCPFC's ones.
- For the Southern Bluefin Tuna, only data from the the Commission for the Conservation of Southern Bluefin Tuna (CCSBT) was kept (i.e. data from the other RFMOs for SBF was removed).
- Original code lists have been mapped with standard FAO code lists (for gears, species, flags). These mappings have been done in collaboration with the RFMOs secretariats.

More details on the processes are provided in the lineage section.

Caracteristics, specific issues and problems that need to be taken into account :

115 views | 121 downloads

See more details...

Indexed in OpenAIRE

Publication date:
February 1, 2018

DOI:
DOI 10.5281/zenodo.1164128

Keyword(s):
tuna | fisheries | tuna atlas | catch | time series | geographic information system | Log school

Related identifiers:
Identical to
https://tunaatlas.d4science.org/geonetwork/srv/eng/catalog.search#/metadata/global_catch_tunaatlasIRD_level2
https://tunaatlas.d4science.org/tunaatlas/index.html?dataset=global_catch_tunaatlasIRD_level2

Communities:
Fisheries and aquaculture

License (for files):
Creative Commons Attribution Non Commercial 4.0 International

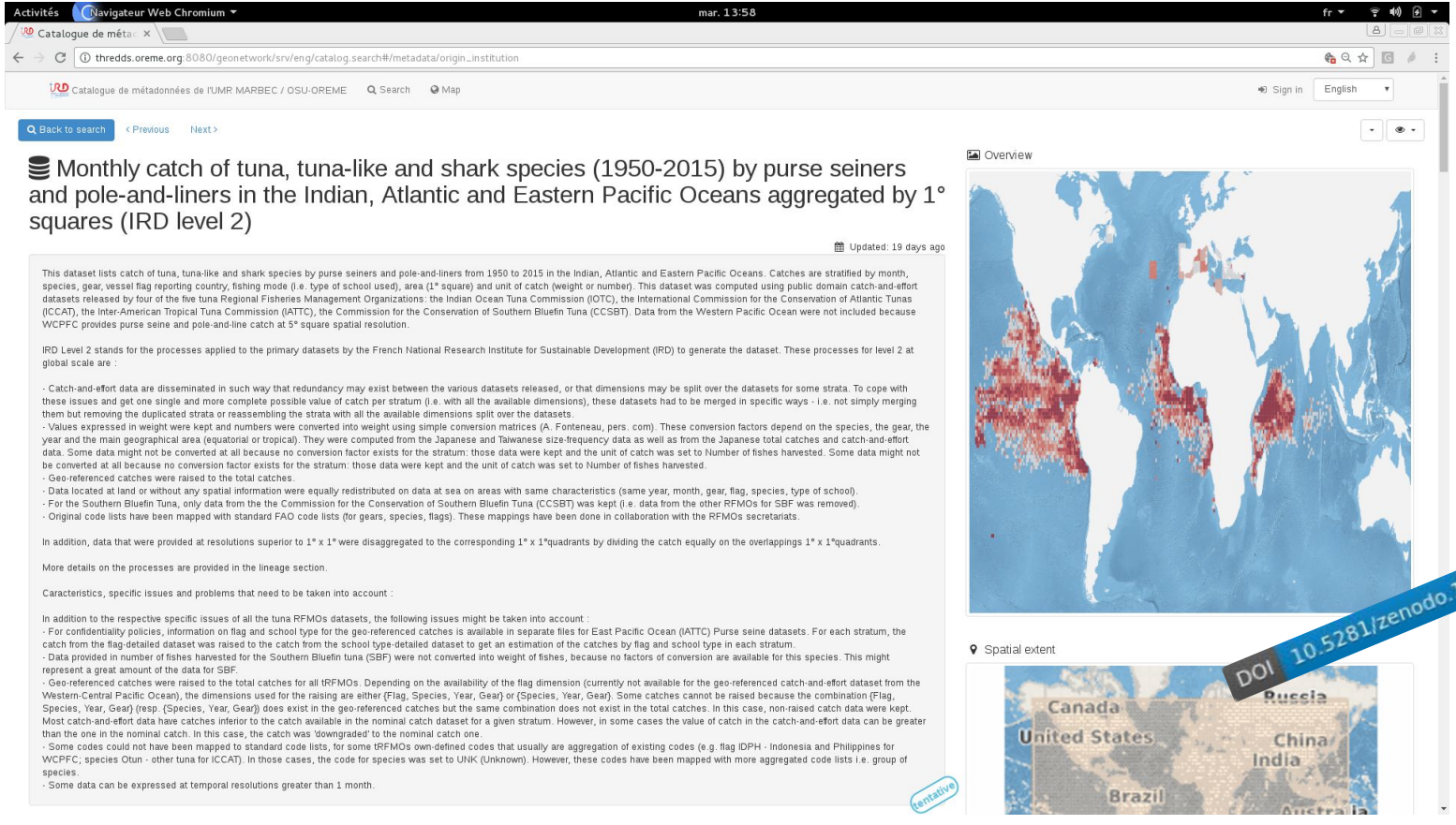
Versions

Version 1 — Feb 1, 2018
10.5281/zenodo.1164128

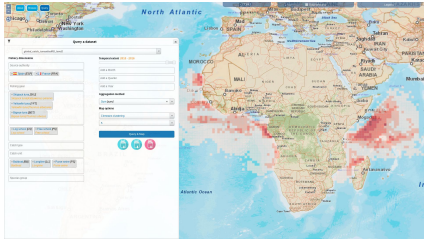
Cite all versions? You can cite all versions by using the DOI

… & Dataverse IRD ??

geoflow (Postgis, GN, GS) => Tuna Fisheries



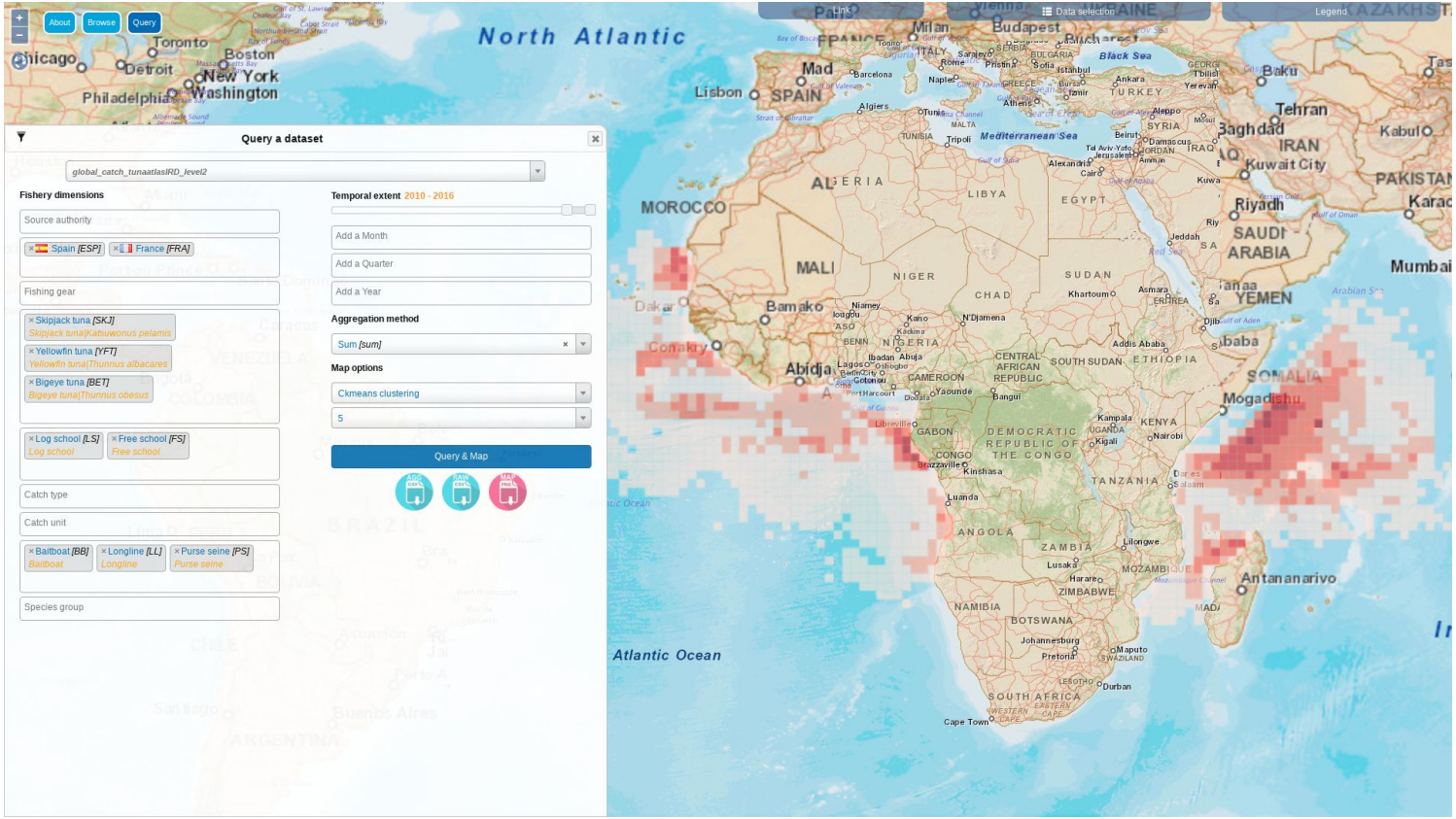
geoflow: ultimate workflow



Assurer les principaux services à partir de “métadonnées riches”:

- Production de métadonnées conformes aux standards OGC data & services
  - 19115 (modèle pivot)
  - 19110 (autre modèle pivot dans un CSV / google doc)
  - 19119 (publiés dans Geoserver)
- Validation du contenu des métadonnées par un viewer FAIR
  - Métadonnées avec DOI,
  - Accès aux données,
  - Visualisation cartographique,
  - Requêtage / Filtres sur les données sur des dimensions ou variables communes à différentes structures de données.

geoflow: ultimate workflow, exemple du Tuna atlas



geoflow: esprit tuning

L’approche par défaut peut être customisée:

- Changer le modèle pivot / format des métadonnées
  - Changer les handlers
- Le format des sources:
  - Serveurs OPeNDAP (eg Thredds): migration des anciens codes,
  - No SQL (MongoDB, Cassandra..), Triple stores / SPARQL
  - Photos géoréférencées,
- Nouvelles actions:
  - migration des anciens codes:
    - publier dans Dataverse avec l’API SWORD,
    - Geonode si résolution CSW-T
  - Autres besoins à analyser..

Conclusion

- Travail en cours, version 0.1
- Objectifs FAIR en privilégiant les standards OGC,
- Approche générique basée sur les métadonnées:
  - Sur la base de simples tableaux (pour n'importe quels jeux de données, bases de données..)
    - modèle pivot (DCMI) et mapping (19115, EML…),
    - Data Structure (19110, CF-conventions): pour CSV, shapefiles, SQL...,
    - Contacts
- Gestion des logs, exécution en ligne...vers la réplicabilité
- Upgrade à suivre dans le cadre de Blue Cloud:
  - Workflow générique (fichiers plats, Postgres..) qui s'appuiera sur un modèle générique pour décrire la structure des données (utilisé par geometa pour produire du 19110)
  - Thredds
  - IPT ?
- Financement Blue Cloud et lien avec la RDA (FAIR, VRE, DMP..)

Perspectives

- Appuyer l'interopérabilité sémantique: interrogation de vocabulaires (SKOS..), registres..
- Validation & Publication des métadonnées INSPIRE
- Gestion de (méta-)données raster
- Gérer plus facilement les contacts avec des annuaires: LDAP & ORCID
- EML & Darwin Core vers OGC
  - LIre les archives sur un IPT
  - Mapping direct ou indirect (via modèle pivot)
  - Produire une version 19110 du Darwin Core
- Interagir avec des plateformes cloud (eg nextcloud, owncloud): eg B2DROP,
- Organiser un nouveau workshop dans le cadre du projet H2020 Blue Cloud / EOSC avec des VRE D4Science (RStudio serveur, Jupyter Notebook)
- Alimenter d'autres documents "officiels" : PGD, Outils de suivi qualité ...

Remerciements

- Blue Cloud H2020 project (grant agreement ID: 862409, Funded under: H2020-EU.3.2.5.1.),
- BlueBridge H2020 project (grant No. 675680): e-infrastructure for marine domain
- OpenAIRE-Connect H2020 project (grant No. 731011): Open Science, Fisheries
- RDA:
  - RDA EU 4.0
  - Fisheries Data Interoperability WG