#### Etat des lieux du moissonnage

de geOrchestra par data.gouv.fr

Landry BREUIL / CRAIG / Geocom 2024

# Data.gouv?

- Data.gouv.fr = plateforme opendata etalab, moteur uData
- Actuellement 4571 organisations et 38034 fiches
- On peut faire ses fiches de MD directement dedans
- Et pointer vers des ressources externes (ou upload)
- Peut moissonner du format DCAT (=XML)
- Guide de publication

#### Tour d'horizon

- DGE: 172 (data4citizen)
- OpenIG:183 (CKAN)
- Arnia/BFC: 5 (onegeo)
- Cigal: 1 (OCS 2008, 12/22)
- PIGMA: 1 (plu 2017)
- CRIGE-PACA, CRAIG: 0
- Datara:0 (prodige)

- IdF: 1014 (ODS)
- Région GE: 301 (d4c)
- PACA: 108 (onegeo)
- Bretagne: 79 (ODS)
- AURA: 11 (transport.data?)
- BFC: 38 (onegeo bfc?)
- IGN: 40 (direct?)
- Rennes: 565 (ods) (+22)

# Un peu d'historique

- RIP Geo.data.gouv.fr
- GN3 avait un endpoint DCAT (srv/XXX/rdf.search)
- GN4 ne l'a plus (perdu dans la grande migration a ES)
- Puis le support de DCAT est revenu en 4.2.(8?)

### Une plateforme de test

- «support de la consommation dcat via csw dans data.gouv »
- https://demo.georchestra.org/geonetwork/ en 4.2.8
- Moissonne les catalogues de la communauté geOrchestra
- Dev G2F 3887, GE 5641, indigeo 1032, CRAIG 533, ...
- Microservice ogc-api-records
- 19565 MD (CSW 18693, OAPIR 11390)
- https://demo.data.gouv.fr
- Configuré moissonneurs csw-dcat et ogc-api-records

#### Premiers résultats

- Csw-dcat:
- 992 archivés
- (archivés = moissonnés a un instant T puis plus présent à T+1)
- 77 échec
- 1923 finis
- 1860 JDD

- Ogc-api-records:
- 303 ignorés (missing id)
- 211 archivés
- 78 échec
- Invalid URL http://ogc.geo-ide.application.i2/
- 1618 finis
- 1533 JDD

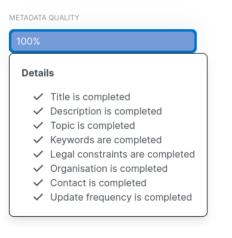


#### **Takeouts**

- Ogc-api-records ne supporte pas la pagination
- On peut demander 15k records mais > 4mn et 82Mo de XML
- proxy\_{read,send}\_timeout 300
- Limite de 2k records sur la démo
- Meme subset ? Aucune idée
- Tout apparaît comme venant du moissoneur configuré

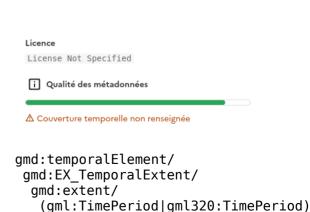
# **Exemples**

- Surface non artificialisée (DH,demo data.gouv)
- Campagnols (DH, GN GeoB, Demo.geor, demo data.gouv)
- Transports Vichy (DH, demo data.gouv)









# Indicateurs de qualité

- Description des données non renseignée
- Documentation des fichiers manquante
- Fréquence de mise à jour non renseignée
- Couverture temporelle non renseignée
- Couverture spatiale non renseignée
- Formats de fichiers non standards
- Licence (a noter que 'License Not Specified' coche en vert?)
- Disponibilité (hits sur ressource?)
- A jour -> en fct de la fréquence ?

#### Améliorations en cours

- Travail de Jean puis ticket puis PR GN
- Support couverture spatiale dans uData
- mapping de la licence (accessConstraints=>dct:license)
- lien vers source d'origine de la MD -> dcat:landingPage
- Mais quid des catalogues moissonnés ?
- affichage des 'autres contacts' non prévu ?
- ils sont dans foaf:Organization et dct:publisher mais sans rôle
- rdf:Description renvoie vers la vision HTML/XML de la MD
- dct:abstract rempli depuis le résumé de la MD iso

#### Process itératif

- faire un pt d'entrée/csw virtuel avec une seule MD dans GN4
- Filter uuid: \$uuid
- créer un harvester csw-dcat sur demo.data.gouv.fr
- modifier/publier sa MD / tpf-rdf.xsl + restart GN
- tester le rendu DCAT
- SERVICE=CSW&REQUEST=GetRecordById&outputSchema=http://www.w3.org/ns/dcat%23&id=\$uuid
- aller dans l'admin demo.data.gouv pour son harvester
- éditer puis prévisualiser le harvester = hit sur la MD (pas auto..)
- => La fiche parfaite (+DH veut useLimitation ?)

# Conclusion

GeoDCAT-AP en 4.4 ?

- Expert XSLT recherché
- Tout est dans tpl-rdf.xsl
- Des questions ?