

Open Data : Les plateformes

19 Octobre 2022 – Séance n°2



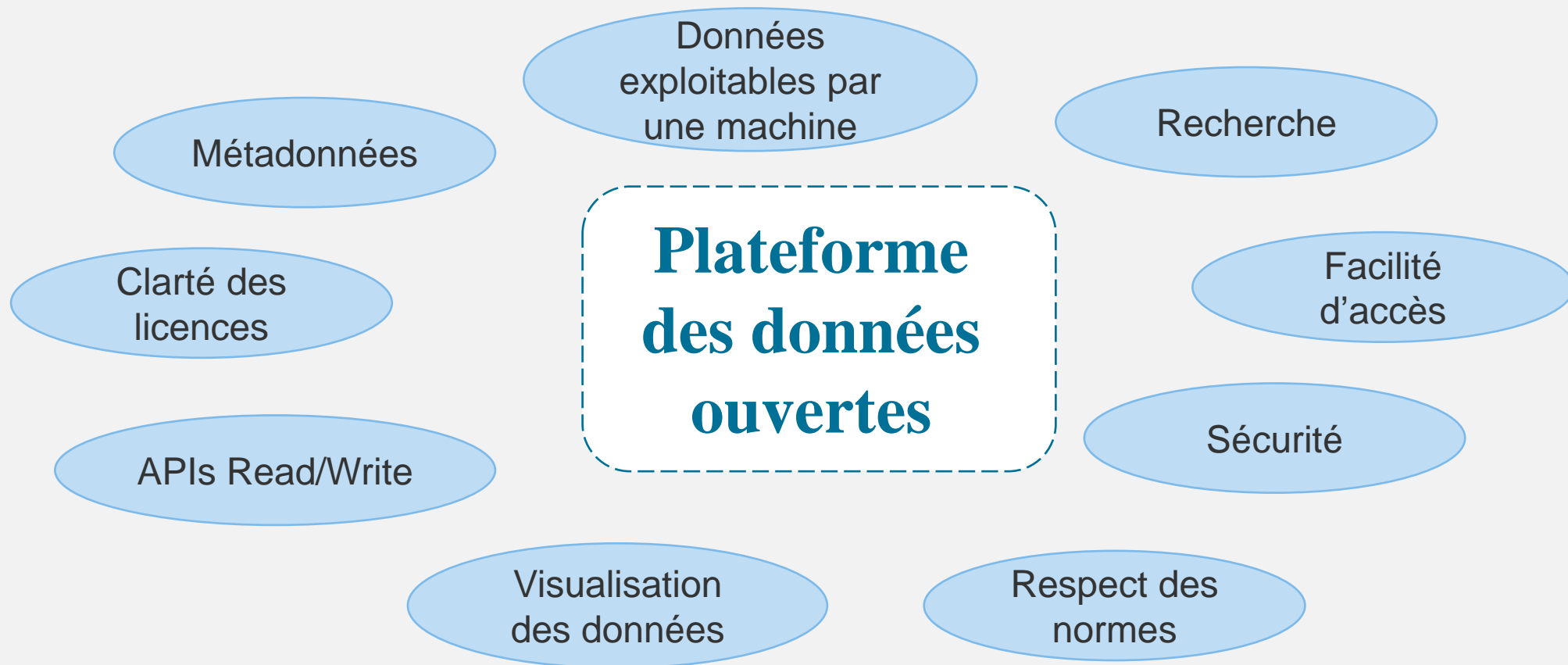
UNITED NATIONS

الاسواق
ESCWA

Shared Prosperity **Dignified Life**



Les plateformes Open Data



Les plateformes Open Data

Deux termes ?

- Catalogue des données ouvertes,
- Plateforme/Portail des données ouvertes,

Trois modèles ?

- Initiative centralisée (au niveau national),
- Initiative décentralisée,
- Initiative fédérée,

Les plateformes Open Data

Une seule plateforme nationale :

Une plate-forme nationale - initiative centralisée

(+)

Un seul et unique point d'entrée aux données : Les jeux de données sont lisibles au niveau national (données et métadonnées)

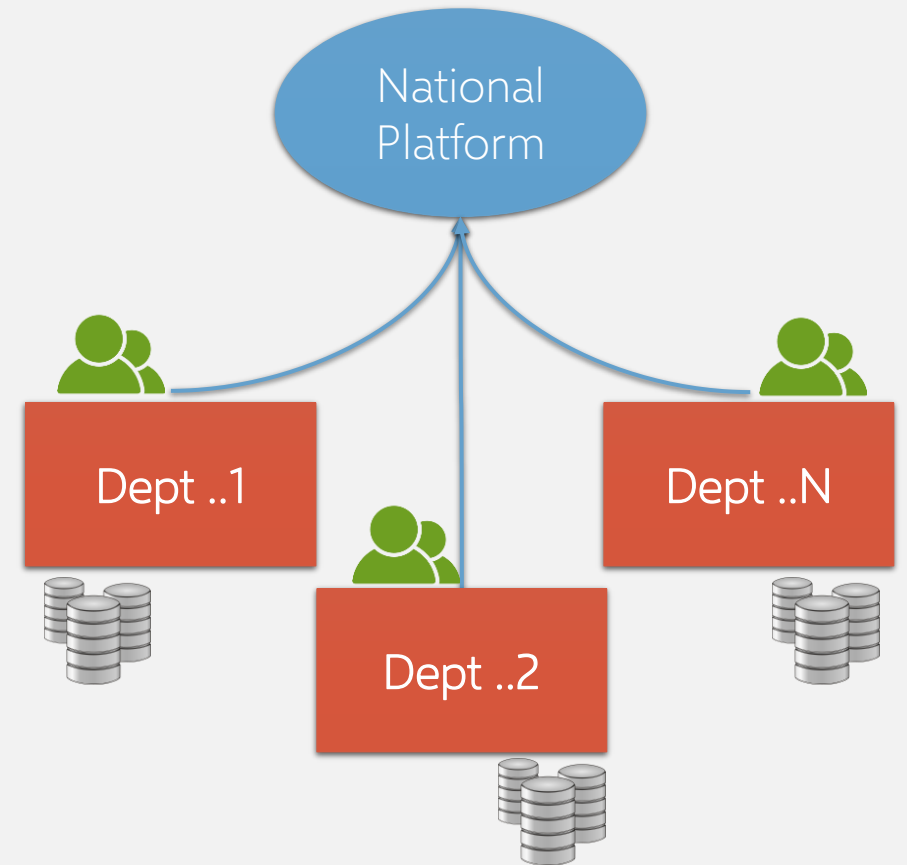
Une seule technologie / Une seule Documentation / une seule API

Facilite la recherche et la réutilisation

(-)

Difficulté de prise en charge des spécificités des autres secteurs

Ne favorise pas l'innovation technologique



Les plateformes Open Data

Plusieurs plateformes décentralisées :

Plateformes sectorielles distinctes - Décentralisées

(+)

Prise en charge de la spécificité sectorielle

Favoriser l'innovation technologique

Favoriser la compétitivité inter-secteur

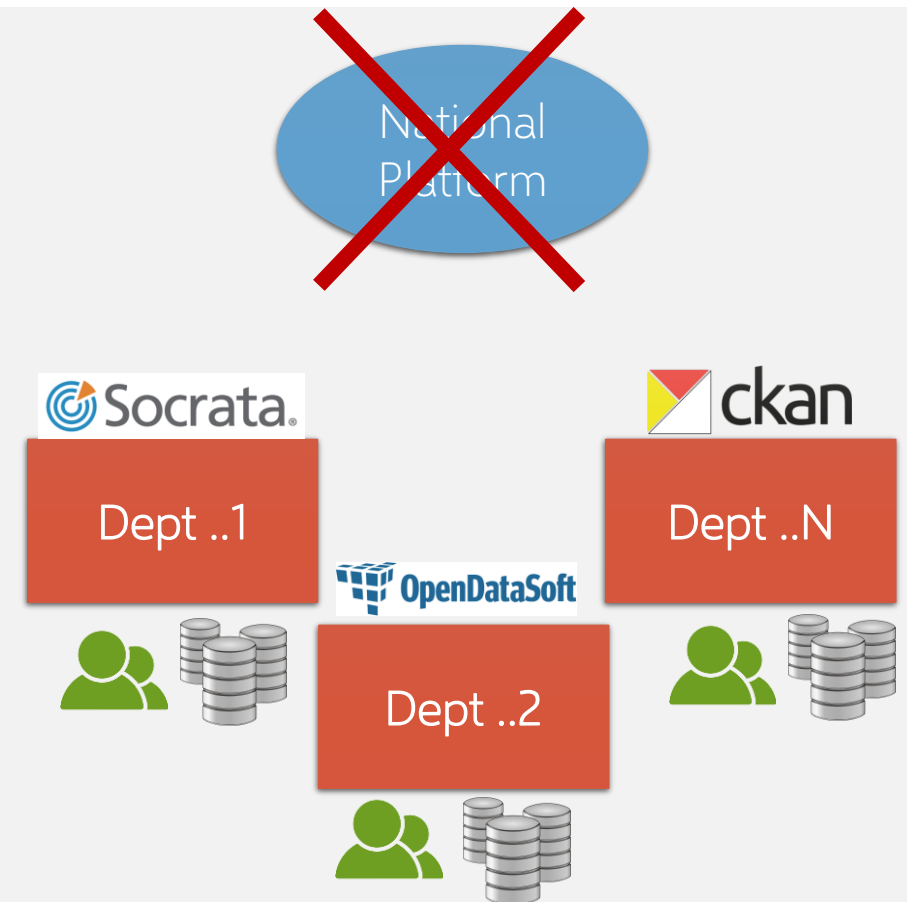
(-)

Absence d'une vision unifiée de l'initiative nationale

Difficulté d'accès et de découverte des données

Des Technologies diversifiées

Plusieurs documentations & plusieurs APIs



Les plateformes Open Data

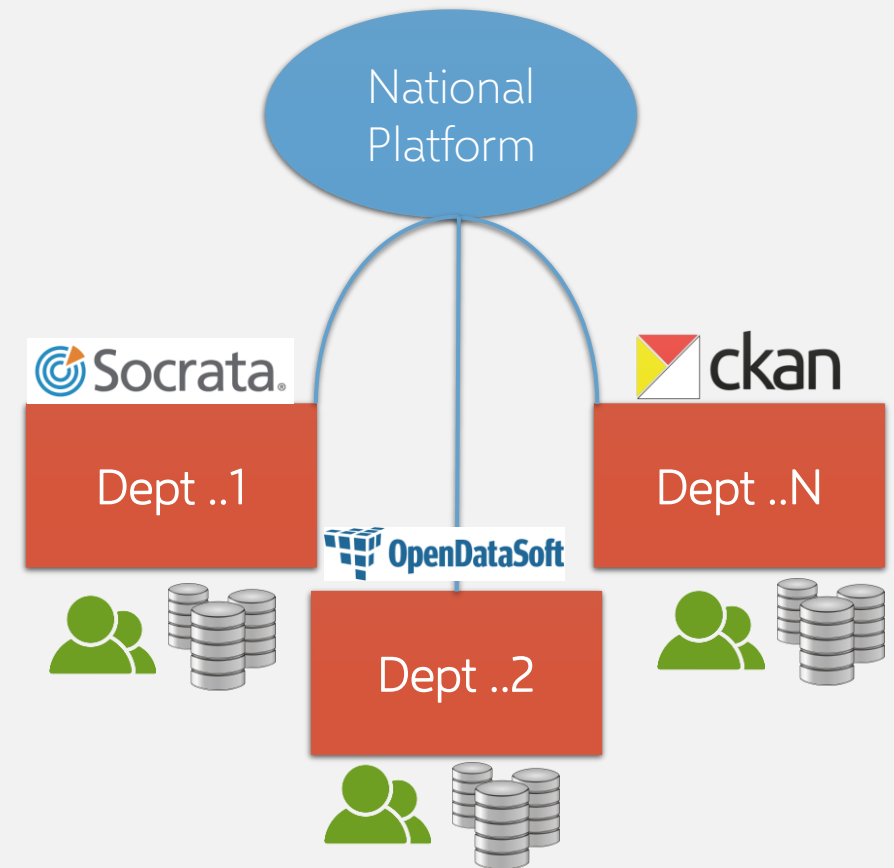
Approche fédérée

Une plateforme nationale fédérée et plusieurs plateformes sectorielles et locales (voir même des plateformes citoyennes).

Prise en charge des spécificités des producteurs de données

Nécessité de la mise en place d'un model unifié des métadonnées

Nécessité de la mise en place des mécanismes de moissonnage pour le référencement des jeux de données au niveau de la plateforme nationale



Les plateformes Open Data

Approche fédérée :

The official portal for European data

173 **36** **1 508 208**

Catalogues Countries Datasets

Trending datasets ?

-  Consolidated list of persons, groups and entities subject to EU financial sanctions
-  Taxpayer Identification Number (TIN)
-  Cosmetic ingredient database (Cosing) - Ingredients and Fragrance inventory
-  BEEG processing statistics
-  The European Commission's Oil Bulletin

Search datasets




Agriculture, Fisheries,
Forestry & Foods



Economy & Finance



Education, Culture &
Sport



Energy



Environment



Government & Public
Sector



Health



International Issues



Justice, Legal System
& Public Safety



Population & Society



Regions & Cities



Science & Technology



Transport

> DATA CATALOGUES

> ALL DATASETS

> EU INSTITUTIONS DATASETS

Les plateformes Open Data

Quel catalogue choisir ?

- ✓ Le producteur de données doit disposer des ressources humaines et financières nécessaires pour la mise en place et la gestion d'une plateforme open data,
- ✓ La plateforme doit garantir le référencement sur la plateforme nationale (interconnexion / moissonnage),
- ✓ Il est recommandé d'utiliser une solution open source,
- ✓ La plateforme doit proposer des mécanismes de participation et d'interaction avec les utilisateurs.

Les mécanismes de participation ?

- ✓ Les demandes d'ouverture de nouveaux jeux de données,
- ✓ Fonctionnalité pour signaler un jeu de données,
- ✓ Les commentaires, les discussions → pour chaque jeu de données,
- ✓ Les cas de réutilisation de données ouvertes,
- ✓ Les propositions d'amélioration de la plateforme open data,

CKAN

CKAN est le catalogue de données open source le plus utilisé par les gouvernements.

Cette solution peut être installée sur un serveur Linux (ou bien en mode container).

Le catalogue CKAN est conçu pour publier et gérer des données au moyen d'une interface IHM ou bien via une API (read/**write**).

Code en Python pour le back-end et de JavaScript pour la partie interface.

La mise en place d'un catalogue de données ouvertes à base de CKAN peut être rapide mais la personnalisation et l'optimisation demandent plus d'investissement qu'une solution clé en main.

Existence d'une communauté de développeurs conséquente , qui assure une évolution constante de la technologie et des outils.

En dehors des outils de recherche, de visualisation, de cartographie classique, et d'interrogation API → CKAN propose aussi des extensions externes accessibles sur ce site : <https://github.com/ckan/ckan/wiki/List-of-extensions>

DKAN

DKAN est une variante de CKAN qui repose sur DRUPAL , un système de gestion de contenu très populaire écrit en PHP au lieu de Python.

DRUPAL, possède sa propre architecture modulaire, avec des milliers de modules disponibles en téléchargement. Ce logiciel permet également de personnaliser les modules et possède une grande communauté de développeurs.

uDATA

uDATA c'est la plateforme développée par EtatLab. Elle est développée en Python et elle est facilement adaptable.

Pour faire tourner une plateforme uData en mode standalone, vous avez besoin d'un simple serveur linux ou MacOS et de suivre la documentation en ligne.

Il est possible de déployer la plateforme avec Docker (container).

Il faut travailler sur la personnalisation.

Le code source est accessible sur : <https://github.com/opendatateam/udata>

Exemples : <https://www.data.gouv.fr> ; <https://data.public.lu> ; <https://dados.gov.pt>

GeoNode

GeoNode est une plateforme open source pour développer des systèmes d'informations géospatiales (GIS) et pour déployer des infrastructures de données géo-spatiales.

Cette plateforme a été conçue pour être facilement extensible et modifiable, et peut être intégrée dans une autre plateforme existante.

Exemples : <https://geonode.wfp.org/>



Shared Prosperity **Dignified Life**



Le catalogue CKAN

CKAN Architecture

Trois (03) types d'installation :

- Installation de CKAN à partir du package
- Installation de CKAN à partir des sources
- Installation de CKAN avec Docker Compose

CKAN Architecture

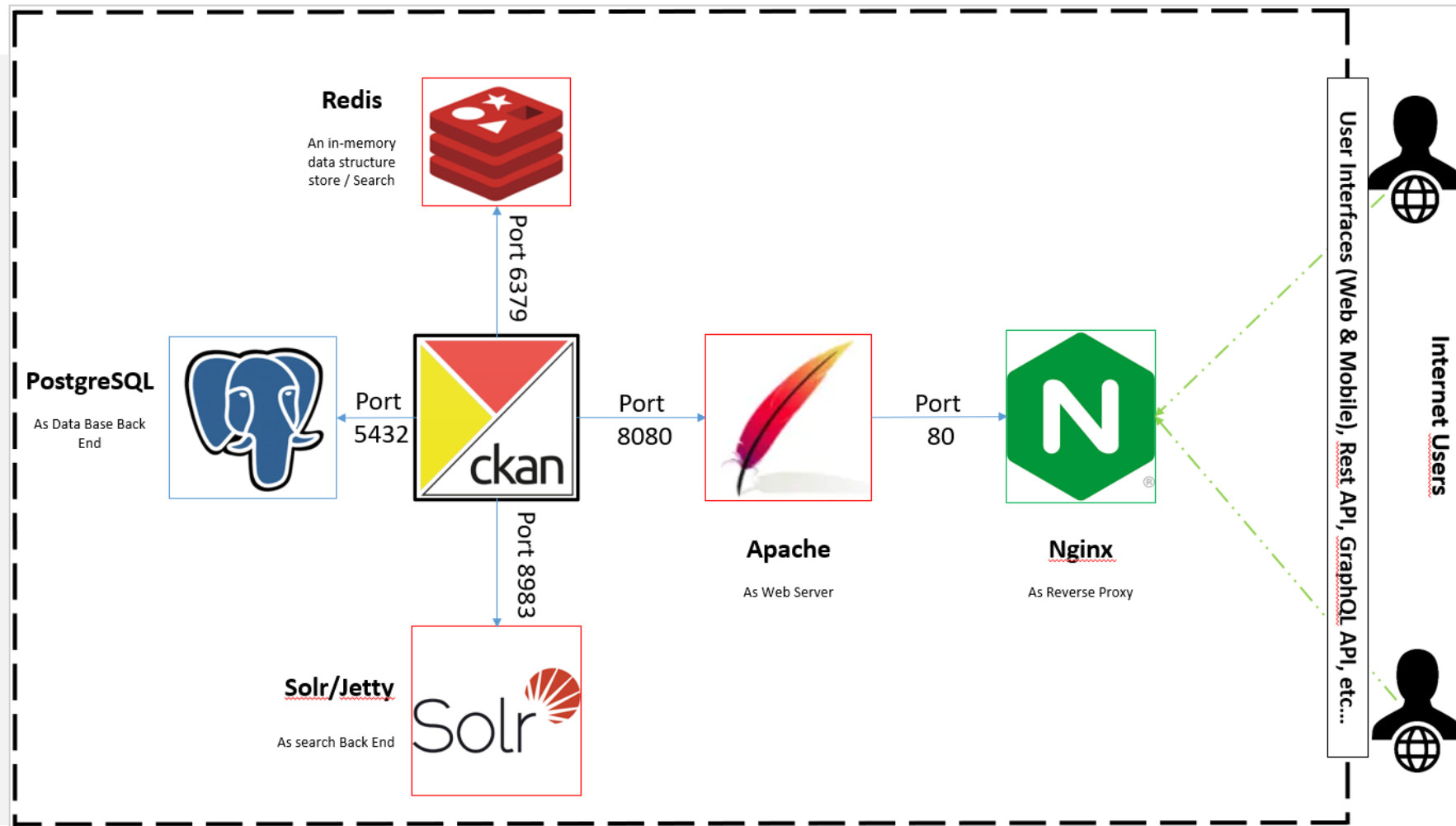
Pour un portail au niveau d'une ville avec des niveaux à faible trafic :

- 2 * Serveurs avec 2 Go de RAM (Web et DB/solr)
- Disque dur de 80 Go sur les deux. (Vous voudrez peut-être plus de stockage sur le serveur Web si vous avez l'intention de faire beaucoup de stockage de fichiers)
- Processeurs double cœur

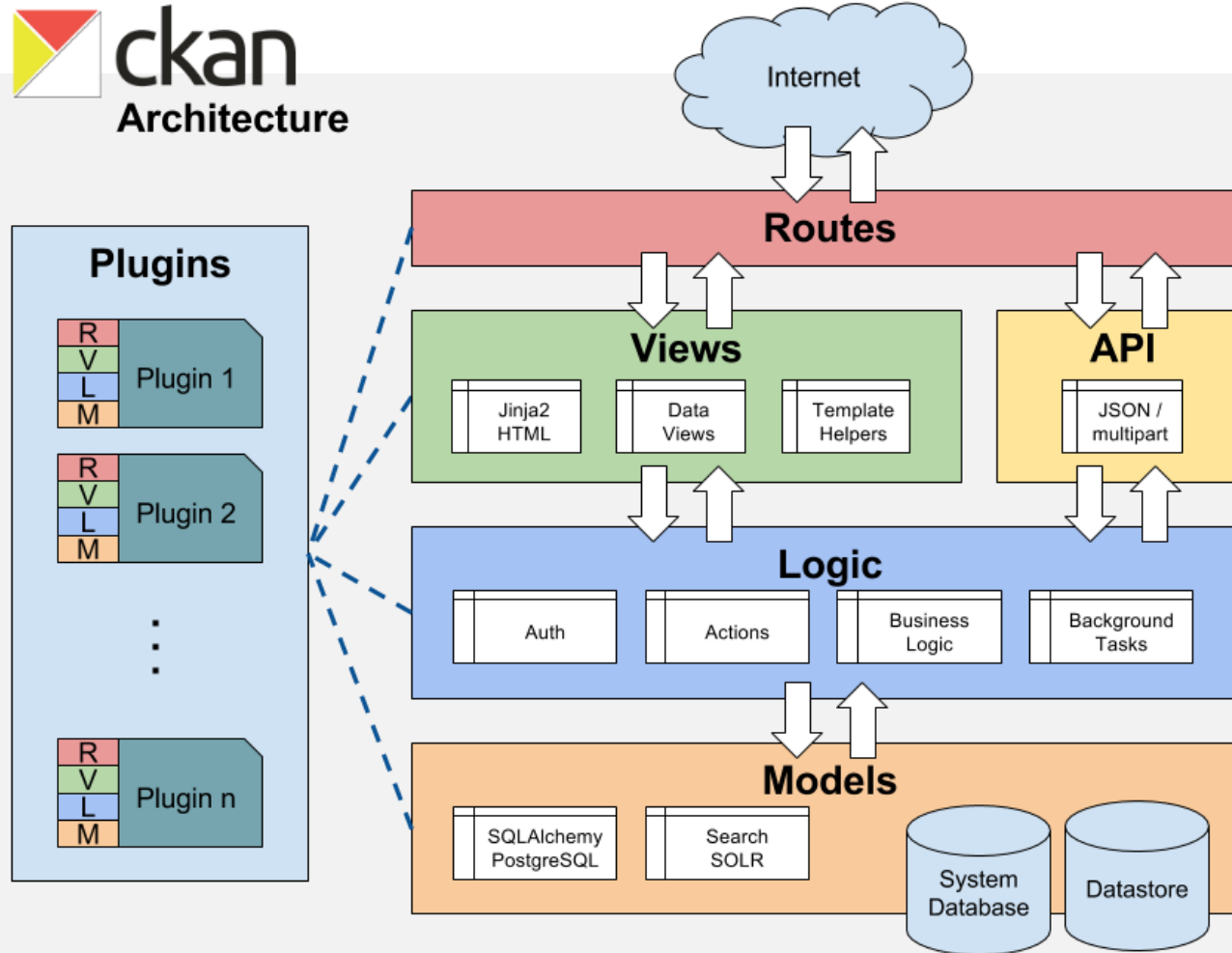
Pour un grand portail national à fort trafic :

- 2 * serveurs avec 8 Go de RAM (un pour le Web et un pour la base de données/solr)
- Disque dur de 160 Go sur les deux. (Vous voudrez peut-être plus de stockage sur le serveur Web si vous avez l'intention de faire beaucoup de stockage de fichiers)
- Processeurs quad core (plus rapide mieux)

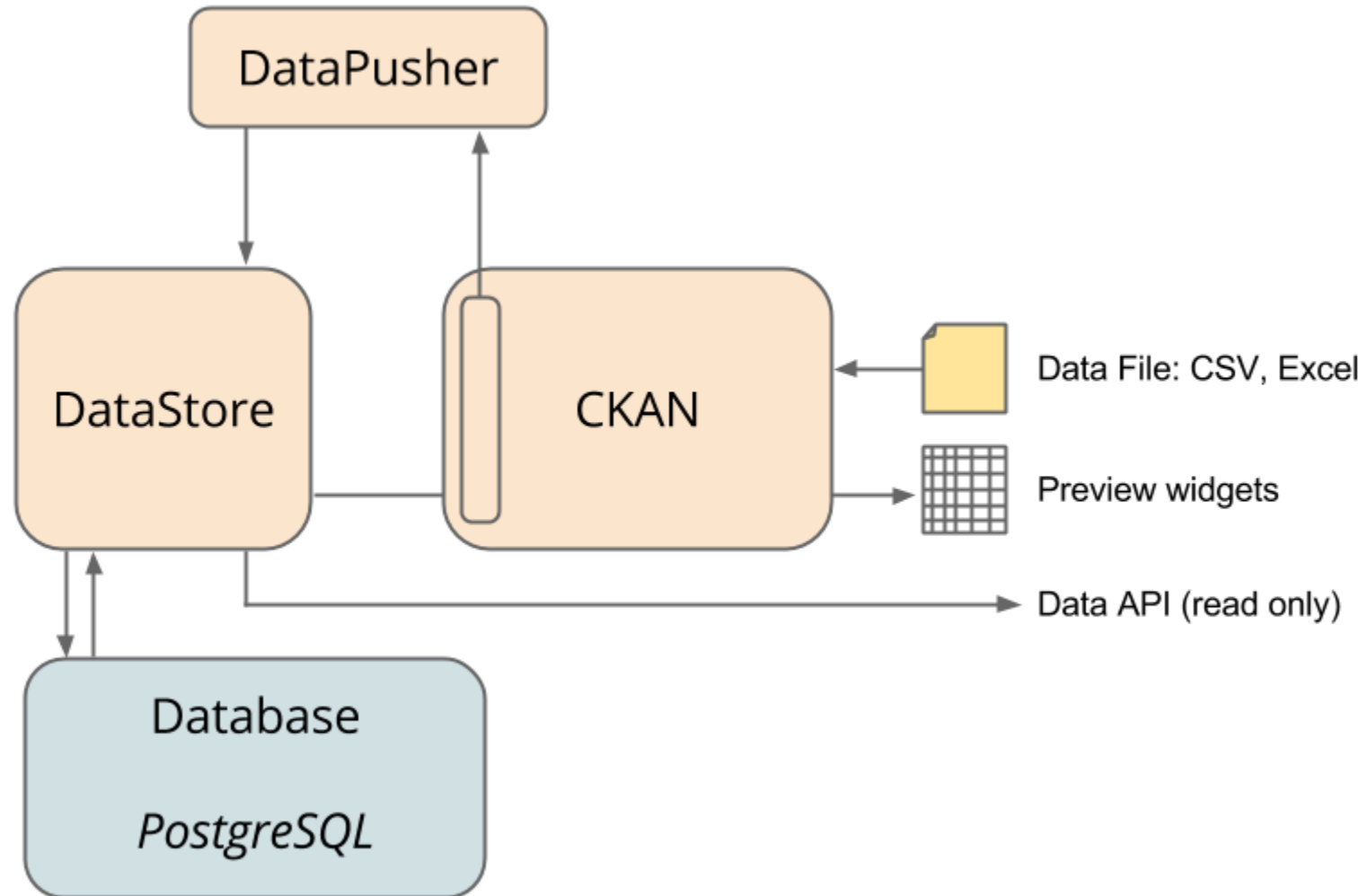
CKAN Architecture



CKAN Architecture



CKAN Architecture



CKAN et les extensions

DataStore + Data Pusher / xloader

L'extension CKAN DataStore fournit une base de données ad hoc pour le stockage de données structurées à partir de distributions d'un jeu de données CKAN. Les données peuvent être extraites des fichiers et stockées dans le DataStore.

Lorsqu'une distribution est ajoutée au DataStore, on obtient :

- Aperçu automatique des données sur la page de la distribution, à l'aide de l'extension Data Explorer
- API Data : permettant de chercher, filtrer et mettre à jour les données, sans télécharger et uploader l'intégralité du fichier de données

Le DataStore est intégré à l'API CKAN et au système d'autorisation.

➔ Le DataStore est généralement utilisé avec le DataPusher, qui téléchargera automatiquement les données vers le DataStore à partir de fichiers appropriés, qu'ils soient téléchargés via le FileStore de CKAN ou via des liens .

CKAN et les extensions

<https://github.com/open-data/ckanext-scheming>

Scheming est une extension permettant de personnaliser des schémas des jeux de données, des groupes et d'organisation dans CKAN.

L'extension des schémas se fait à l'aide de fichiers JSON ou YAML qui incluent des règles de validation, la modification d'extraits de modèle de formulaire et des extraits d'affichage.

CKAN et les extensions

<https://github.com/open-data/ckanext-fluent>

Fluent est une extension permettant d'ajouter du texte et des balises multilingues aux jeux de données, aux groupes et organisations de CKAN.

Fluent fonctionne avec Scheming en fournissant des validateurs et des extraits de code pouvant être utilisés dans les schémas personnalisés.

Fluent affiche des zones d'édition pour toutes les langues pour chaque champ dans le formulaire d'édition, mais par défaut n'affiche que la valeur de la langue de l'utilisateur lors de l'affichage d'un jeu de données.

Toutes les valeurs multilingues sont renvoyées et créées/mises à jour via l'API à l'aide d'un objet JSON.

Par exemple. un champ Fluent "label" peut avoir la valeur :

```
... "label": { "ar": "كتب", "fr": "Livres"}, ...
```

CKAN et les extensions

Les autres extensions :

webpage_view / pdf_view / officedocs_view / basic-charts

Spatial_metadata / spatial_query navigablemap / geojson_view

Harvest / dcat

Pages / googleanalytics / showcase / datarequests / disqus / rating / contact



Shared Prosperity **Dignified Life**



Cas du Royaume du Maroc

Plateforme Open Data

CKAN et les extensions

Une nouvelle plateforme nationale : CKAN
2.9.3 + Drupal 9

Les extensions au niveau de la nouvelle
plateforme nationale

```
▼ "extensions": [  
  "duo",  
  "duo_dataset",  
  "duo_organization",  
  "duo_group",  
  "stats",  
  "text_preview",  
  "recline_preview",  
  "multilingual_group",  
  "multilingual_tag",  
  "datastore",  
  "scheming_groups",  
  "scheming_organizations",  
  "pdf_view",  
  "officedocs_view"  
]
```


Au niveau central , Comment enrichir la nouvelle plateforme?

Offre

Le schéma des métadonnées

Les API pour l'interrogation des données ouvertes

Les mécanismes pour améliorer la qualité des métadonnées (données ?)

Demande

Mécanisme pour la proposition d'ajout de nouveaux jeux de données,

Réutilisation

Valorisation et partage des cas de réutilisation de données,

Mesurer l'impact de la réutilisation des données,

Des canaux de communication entre producteurs & ré-utilisateurs, ainsi que des mécanismes de signalement d'anomalies

Cas du Royaume du Maroc: la Plateforme ?

Au niveau des producteurs des données, **Commencer à publier les premiers jeux de données les plus importants !**

- ❑ **Faible nombre** de jeux de données avec une **faible fréquence** de mise à jour → Insertion manuelle directe sur la plateforme nationale.
- ❑ **+10 jeux de données** avec une **fréquence** de mise à jour **régulière** → Approche automatisée, qui se base sur l'exploitation des APIs (d'écriture) offerts par la plateforme nationale.
- ❑ **Nombre** de jeux de données **importants** avec une **mise à jour fréquente** → Étudier la mise en place d'une plateforme spécifique **avec obligation de respect de la procédure de moissonnage** des données vers la plateforme nationale (même schémas de métadonnées !).



Shared Prosperity **Dignified Life**



Discussion