Tiwara

# Document de présentation V1.0

## Data Team - Alliance Sahel

Ce document a pour objectif de présenter le projet Tiwara implémenté par l'équipe Data de l'AFD.

Ce projet consiste en la mise en place d'un tableau de bord agrégeant différentes sources de données sur la région du Sahel.

La technologie utilisée est R/Shiny.

Les différents points abordés dans ce document seront :

- Les sources de données utilisées

- Les représentations réalisées

- La structure informatique du projet et le code

# 1) Les sources de données

Différentes sources ont été utilisées dans ce projet afin de donner une vision la plus complète possible de la situation au sahel.

* DHS Program
* World Bank
* ACLED
* Inform Risk
* AFD
* UNHCR
* ILO
* Data Financial Index

a) DHS program

Le programme DHS (Demographic and Health Surveys) a pour vocation de collecter des données par des méthodes de sondage sur des sujets de santé, de situation sociale, d'alimentation...

Ces données permettent de mieux comprendre la situation dans les pays concernés a une maille géographique plus fine que le pays.

Des informations supplémentaires sont disponibles sur le site du projet : https://www.dhsprogram.com/Who-We-Are/About-Us.cfm

Dans le cadre du projet Tiwara, les indicateurs exploités sont ceux concernant l'éducation :

- Taux d'alphabétisation

- Taux de scolarisation

- Nombre d'années d'éducation

Les données étant issues de sondages sur différents pays, il est important de noter que les années de recueil diffèrent d'un pays à l'autre.

L'exploitation de cette donnée demande une inscription sur le site du DHS Program ainsi qu'une justification de l'usage.

b) World Bank - World Development Indicators

La World Bank est une organisation visant à développer des solutions durables pour diminuer la pauvreté et accompagner les pays en développement.

Nous utilisons ici l'évolution des taux de scolarisation sur les différents pays afin de pouvoir observer les tendances.

Il s'agit d'open data.

Plus d'informations : https://www.worldbank.org/en/who-we-are

c) ACLED - Armed Conflict Location & Event Data

Le projet ACLED vise à recenser l'ensemble des crises et des conflits en Afrique, Asie du Sud, Asie du Sud Est, Moyen Orient, Europe et Amérique Latine.

On utilise les données issues de l'Acled afin de représenter les situations de crise au Sahel.

Cette information permet de par sa forte précision géographique (données géolocalisées) de détecter des zones sensibles ou fréquemment soumises à des conflits.

Il s'agit d'open data.

Plus d'informations : https://www.acleddata.com/about-acled/

d) Inform Risk

La donnée Inform Risk fournit à un niveau régional une grande quantité d'indicateurs de risque.

On retrouve des risques de santé, de malnutrition, des risques sanitaires, sociaux, éducatifs...

La donnée est ici représentée sous forme de cartes, mais également exploitée pour regrouper les régions entre elles (clustering).

Il s'agit d'open data.

Plus d'informations sur Inform Risk : http://www.inform-index.org/

e) AFD

L'AFD met à disposition des informations sur les projets qu'elle finance.

Cette information permet de trouver une géolocalisation ainsi qu'une description des projets.

Il s'agit d'open data.

Plus d'informations : https://opendata.afd.fr

f) UNHCR - UN Refugees Agency

L'UNHCR fournit des données concernant les flux migratoires et leur évolution dans le temps (donnée annuelle).

Cette information est au niveau des pays. Différents types de flux migratoires sont fournis :

- Migration

- Déplacements internes

- Recherche d'asile

- ...

Il s'agit d'open data.

Plus d'informations : http://popstats.unhcr.org/en/overview

g) ILO - International Labour Organisation

L'ILO est une organisation qui rapproche 187 états membres afin de définir des conditions de travail décentes pour tous.

Les données utilisées concernent les taux d'emploi des jeunes ainsi que leur évolution.

Il s'agit d'open data.

Plus d'informations : https://www.ilo.org/global/about-the-ilo/lang--en/index.htm

h) Data Financial Index (World Bank + Bill and Melinda Gates Fundation)

Cette donnée donne des informations issues de sondages sur la situation financière des individus dans le monde.

Elle contient des indicateurs d'épargne, de prêts, de possessions, de situation salariale...

On l'exploite pour donner un aperçu de la situation financière dans les pays.

Cette donnée demande une inscription ainsi qu'un motif d'exploitation pour être utilisée.

Plus d'informations : http://microdata.worldbank.org/index.php/catalog/3324

# 2) Les représentations des données

L'application est divisée en 3 axes :

- Indicateurs socio-économiques

- Crises et conflits

- Démographie

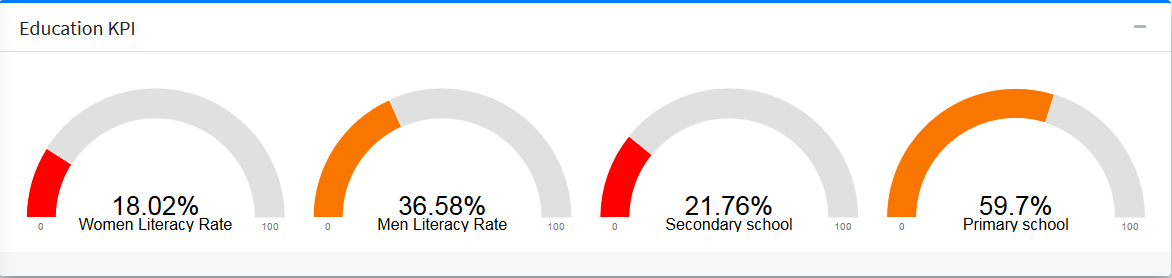
Ces axes sont abordés par l'intermédiaire d'onglets dans lesquels il y a des visualisations des données.

## A) Indicateurs socio-économiques

a) General Outlook

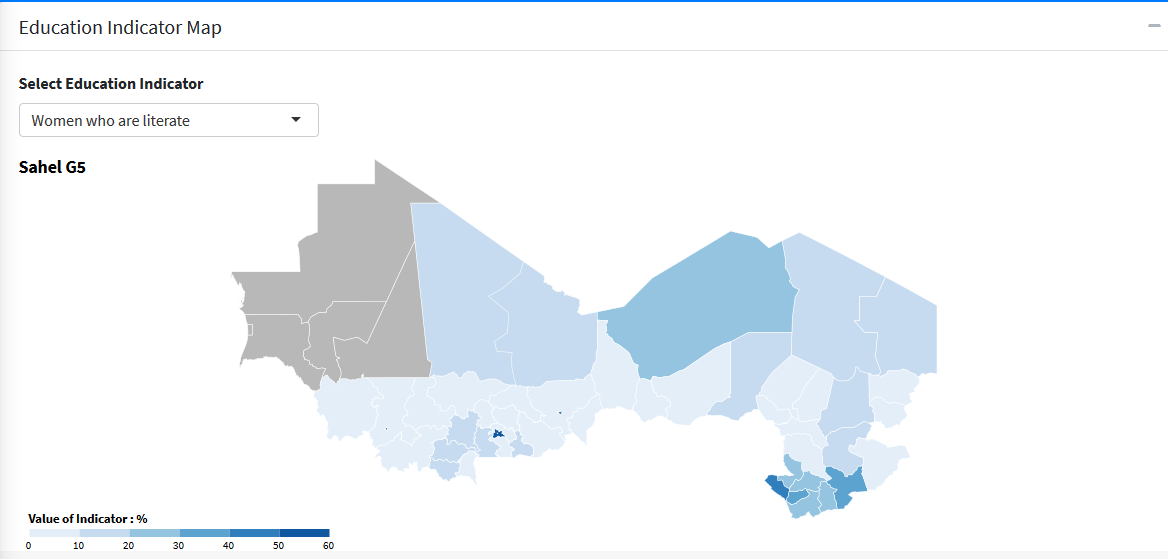
Cet onglet reprend plusieurs indicateurs assez divers. Il est sensible au pays sélectionné(s) dans la barre de gauche.

En haut, on retrouve 4 KPI sous la forme de Gauge Chart.



Ces KPI sont basés sur les données DHS. Les valeurs manquantes ont été supprimées dans le calcul des moyennes.

On trouve en dessous une carte représentant différentes variables liées à l'éducation.



Il est possible de sélectionner la variable que l'on souhaite représenter à l'aide du menu déroulant situé au-dessus de la carte.

Cette carte est basée sur les données DHS au niveau régional.

Point d'attention : Les populations dans les quintiles de richesse sont calculés par rapport au pays de la région, non du sélecteur de la barre de gauche.

En bas de la page, on trouve un graphique en barre représentant les mêmes données que la carte (DHS).

On peut sélectionner la variable à représenter avec le sélecteur du dessus ainsi que le nombre de régions (sélecteur d'en dessous) avec les valeurs les plus hautes.



b) Education

Cet onglet dépend du choix de pays dans la barre de gauche.

En haut de l'onglet Education, on retrouve des indicateurs de ratios élèves/professeurs basés sur les données World Development Indicators (World Bank).

Ces indicateurs dépendent également de la période sélectionnée dans l'encart "Gender Parity Index and School Enrolment (Primary, Secondary & Tertiary) situé juste en dessous.

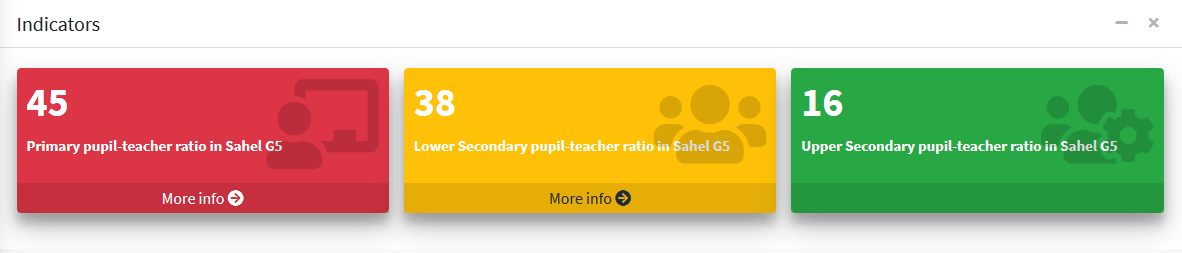
Les valeurs manquantes ne sont pas prises en compte lors du calcul des indicateurs.

La couleur des indicateurs change en fonction de leur valeur :

- Si le ratio est inférieur à 20, l'indicateur est vert

- Si le ratio est entre 20 et 40, l'indicateur est jaune

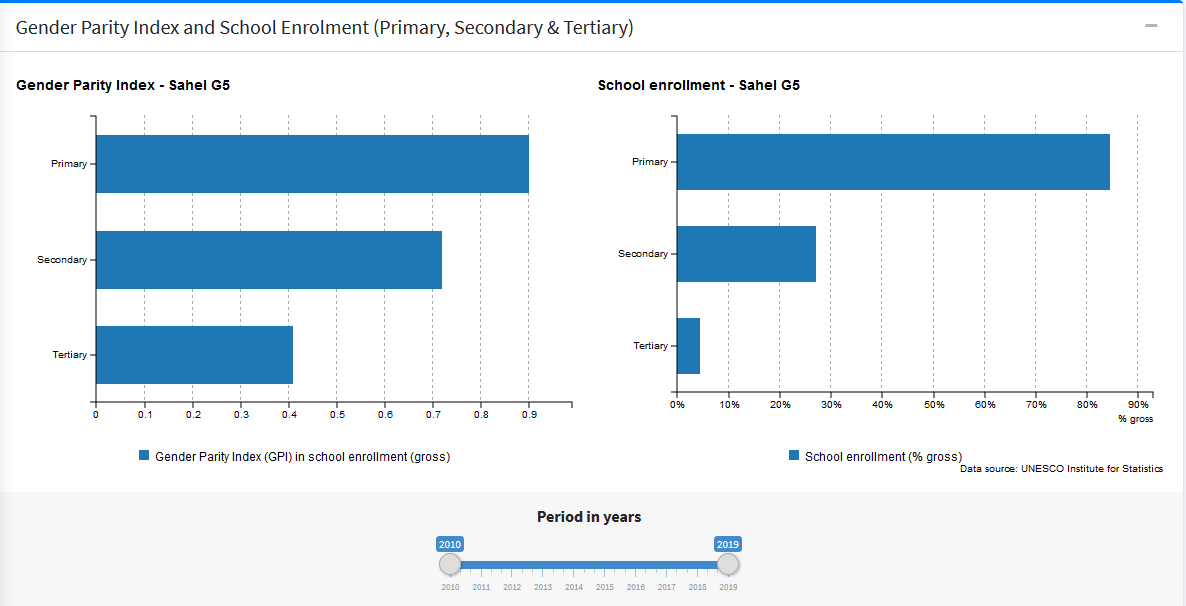
- Si le ratio est supérieur à 40, l'indicateur est rouge



On trouve en dessous deux graphiques en barre basés sur les mêmes données et les mêmes sélecteurs.

Celui de gauche représente l'égalité des sexes dans les différents types d'établissements scolaires.

Celui de droite représente la scolarisation à différents niveaux.



On trouve en dessous un graphique en bulles basé sur les données DHS. Les valeurs manquantes ont été supprimées lors du calcul des moyennes.

La représentation se fait au niveau régional.

La couleur représente le pays.

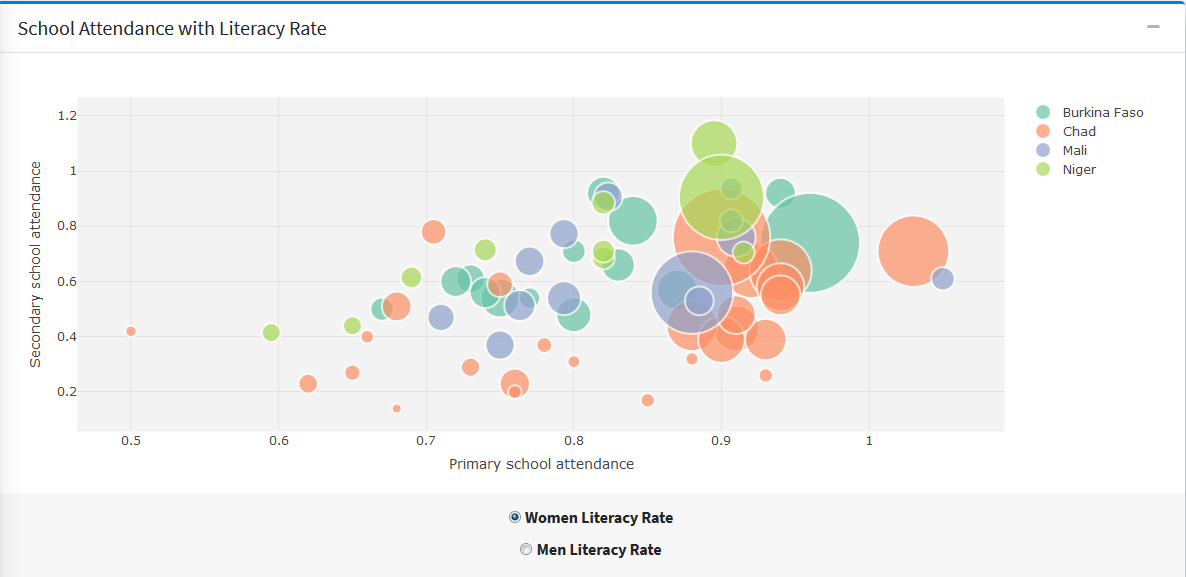
L'axe des abscisses représente la scolarisation au niveau primaire.

L'axe des ordonnées représente la scolarisation au niveau secondaire.

La taille des bulles représente la variable sélectionnée en dessous :

- Taux d'alphabétisation des femmes

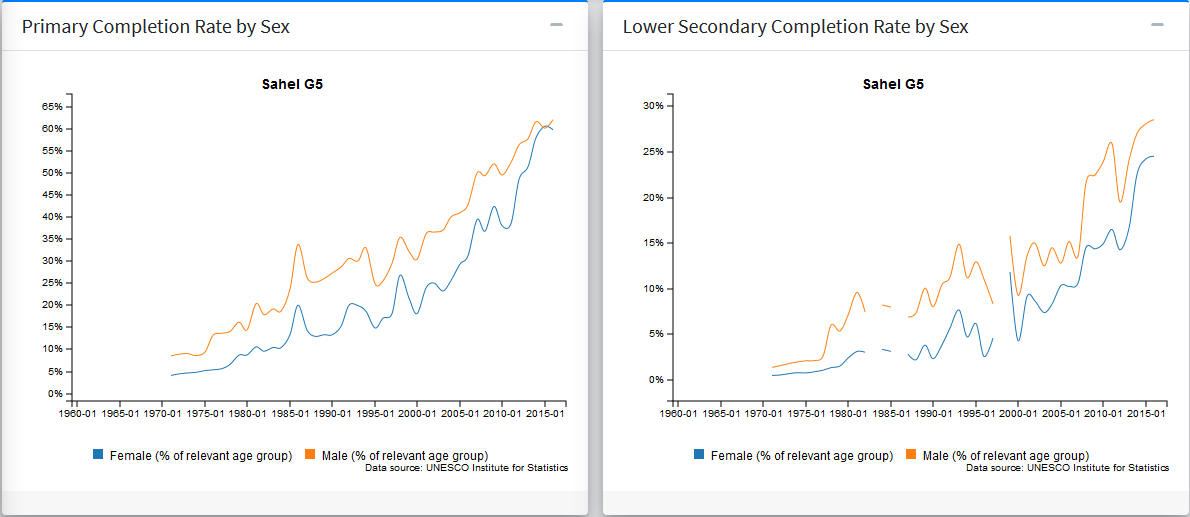
- Taux d'alphabétisation des hommes

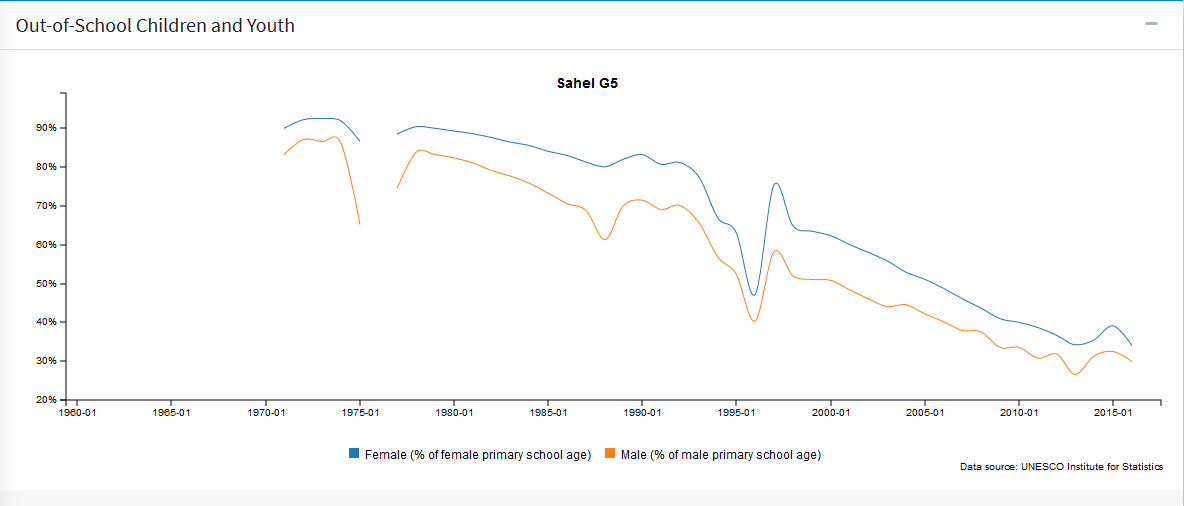


On trouve en dessous deux graphiques montrant l'évolution des sorties de cycles scolaires (primaire, secondaire) par sexe depuis 1960 par genre.

Ces graphiques se basent sur les World Development Indicators. Ils permettent de voir si l'écart entre les genres se réduit ou non.

En dessous on trouve un graphique basé sur les mêmes données représentant la part d'enfants non scolarisés dans le temps par genre.





c) Carto Bar

Cet onglet intègre l'outil développé par Uber Technologies nommé Kepler.gl

Il permet de créer des visualisations cartographiques basées sur des fichiers csv, json, geojson.

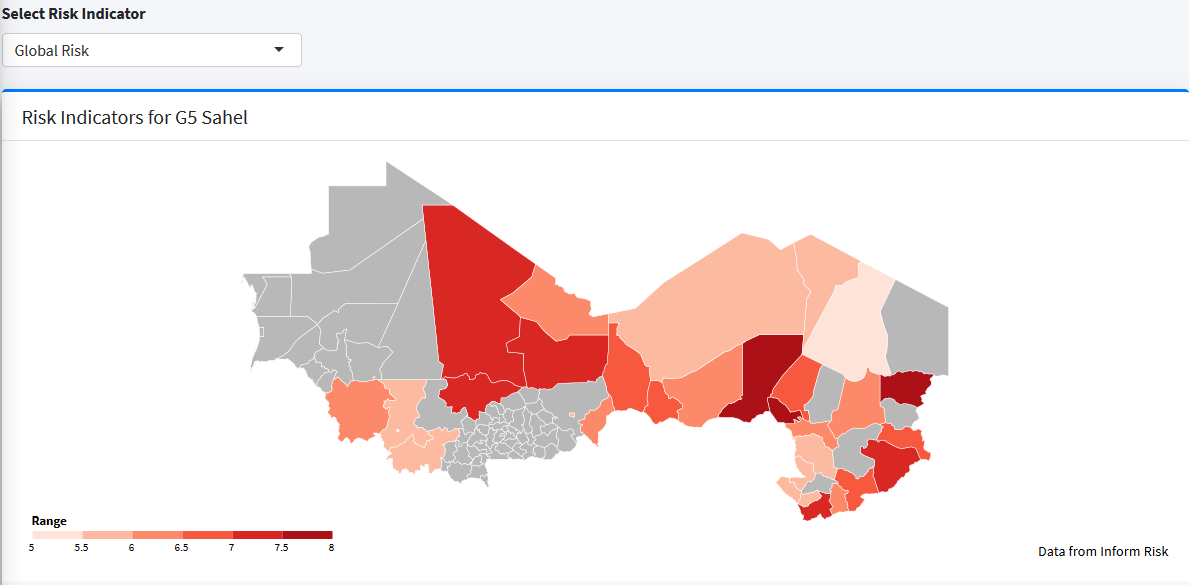
## B) Crises et Conflits

a) INFORM Risk

Cet onglet est sensible au pays sélectionné dans la barre de gauche.

On trouve en haut un sélecteur permettant de choisir l'indicateur de risque à représenter sur la carte située juste en dessous.

Les indicateurs sont au niveau régional et sont issus de la donnée Inform Risk.



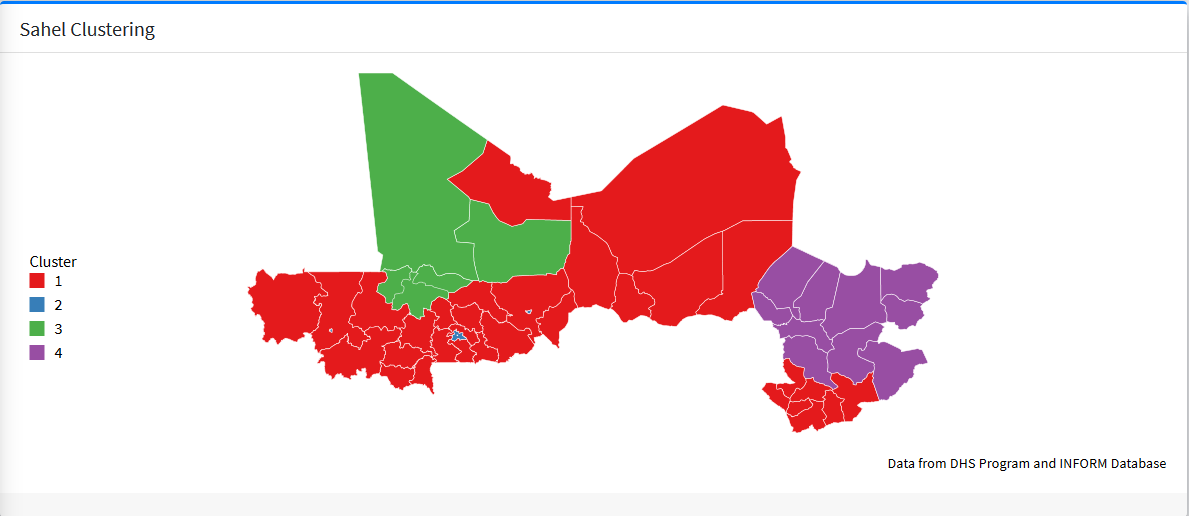
En dessous on retrouve une carte présentant le résultat d'un clustering.

Ce clustering est basé sur les données DHS et INFORM

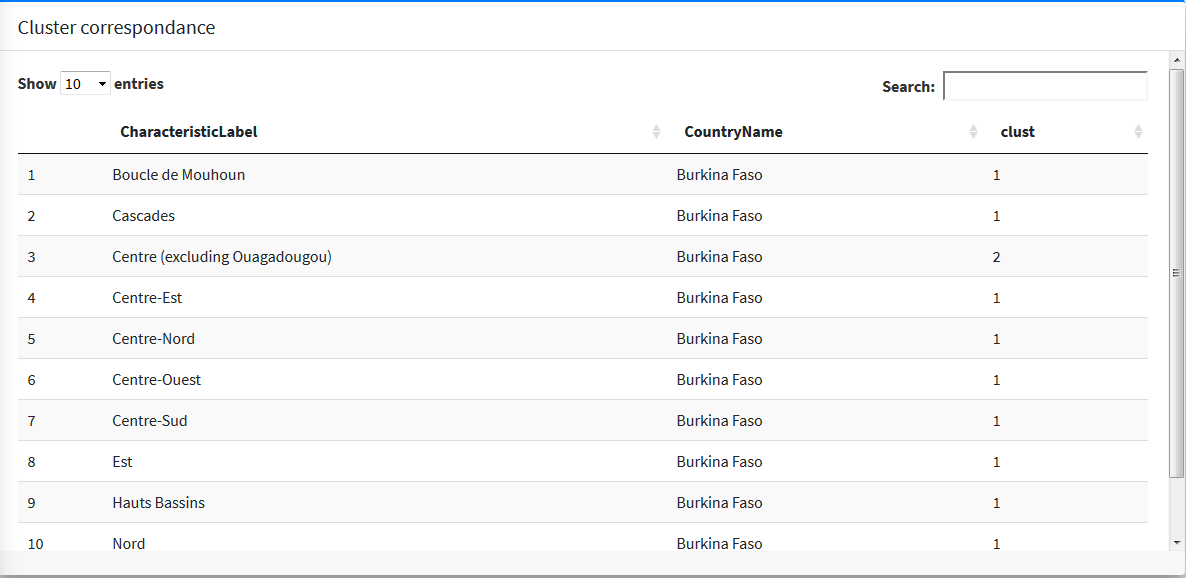
Les données sont jointes au niveau de la région, et on exploite uniquement les données qui sont présentes dans les 2 datasets.

Les valeurs manquantes sont imputées par la moyenne.

Il y a beaucoup de variables fortement corrélées, certaines ont été supprimées pour permettre le clustering.



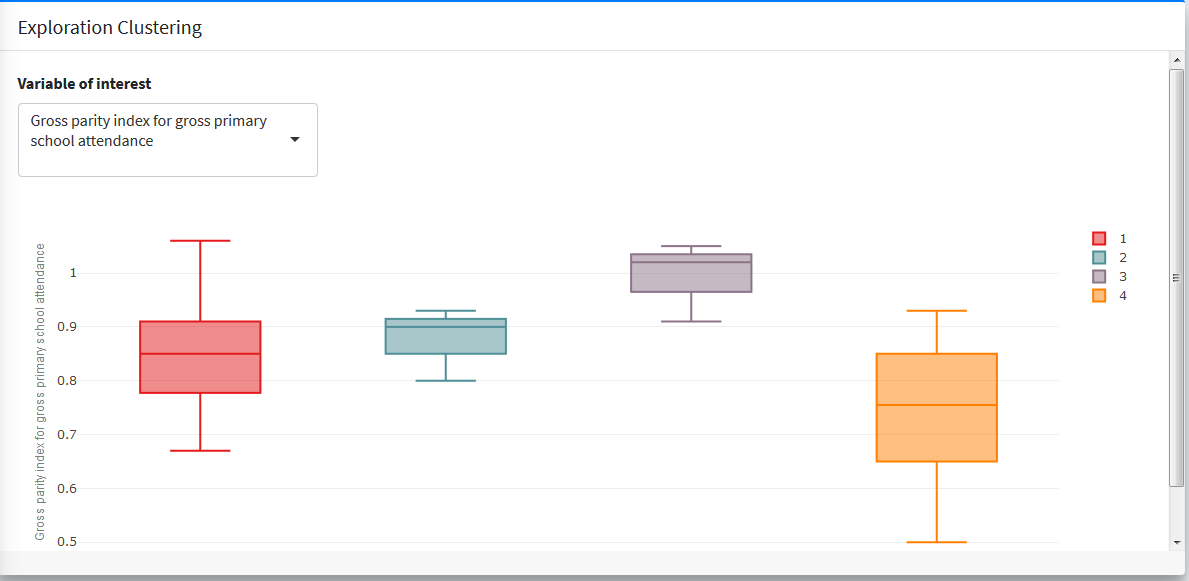
On trouve en dessous une table de données permettant de connaître les correspondances entre les régions et les clusters.



On trouve en dessous des boîtes à moustaches par cluster pour les différentes variables ayant participé au clustering.

Ces boîtes à moustaches permettent de comprendre les profils des clusters.

Le sélecteur permet de choisir la variable à représenter.



b) ACLED Reports

Cet onglet est sensible au pays sélectionné dans la barre de gauche.

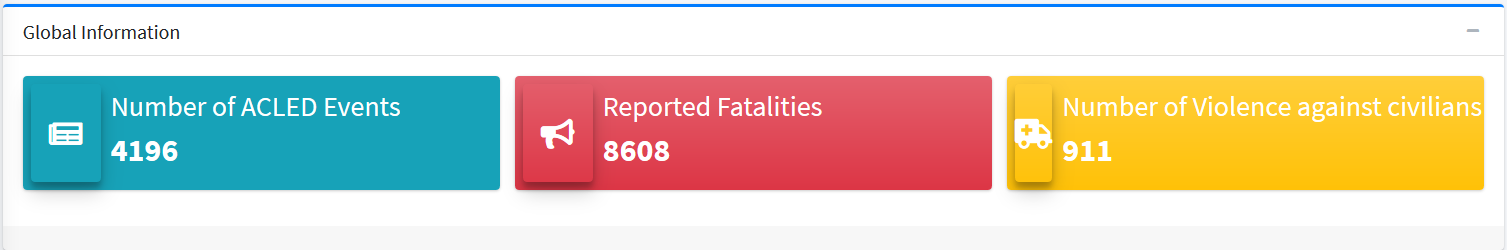
On trouve en haut de la page 3 encarts informatifs sur :

- Le nombre d'événements recensés par l'ACLED

- Le nombre de morts durant ces événements

- Le nombre de violences faites aux civils

Ces valeurs sont basées sur les données ACLED

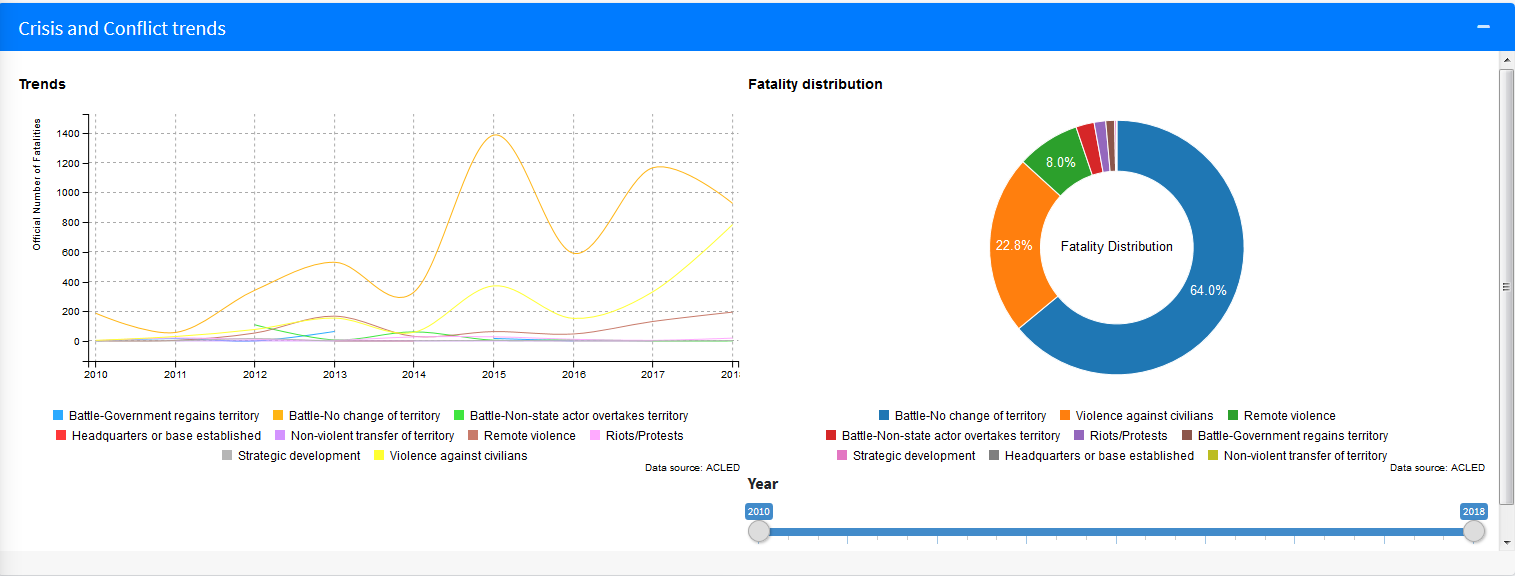


On trouve en dessous deux graphiques :

- Un représentant l'évolution du nombre de morts dans les différents types d'événement recensés depuis 2010.

- L'autre représente pour la période sélectionnée (sélecteur juste en dessous) la répartition des décès par type d'événement.

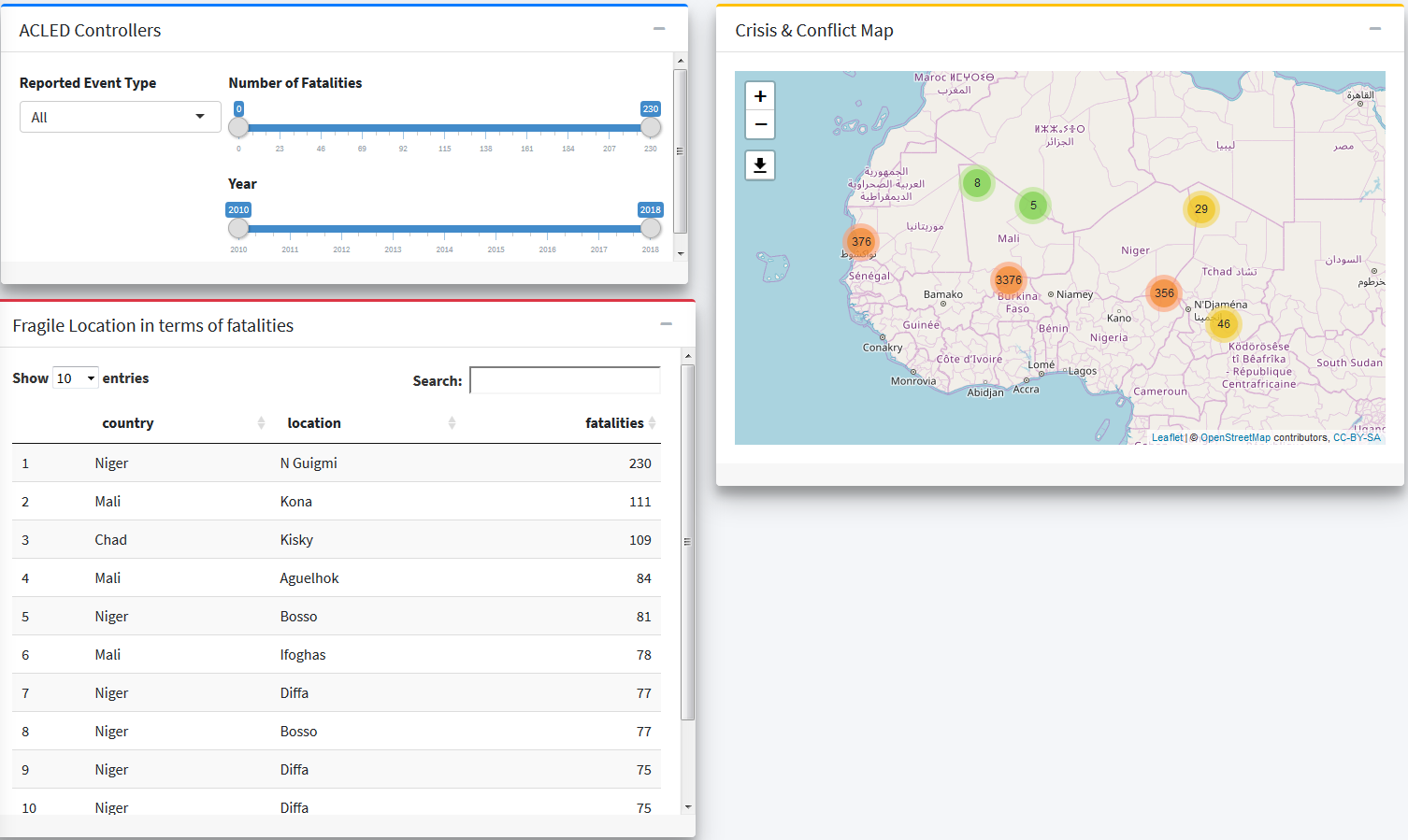
Ces graphiques sont basés sur les données ACLED



On trouve en dessous un ensemble de panneaux permettant de sélectionner des événements (par type, nombre de décès, période) et d'avoir plus d'informations dessus.

Les informations se trouvent dans la table de donnée ainsi que sur la carte.

En cliquant sur une ligne de la table, la carte montre l'événement sélectionné.



Les données proviennent d'ACLED

## C) Démographie

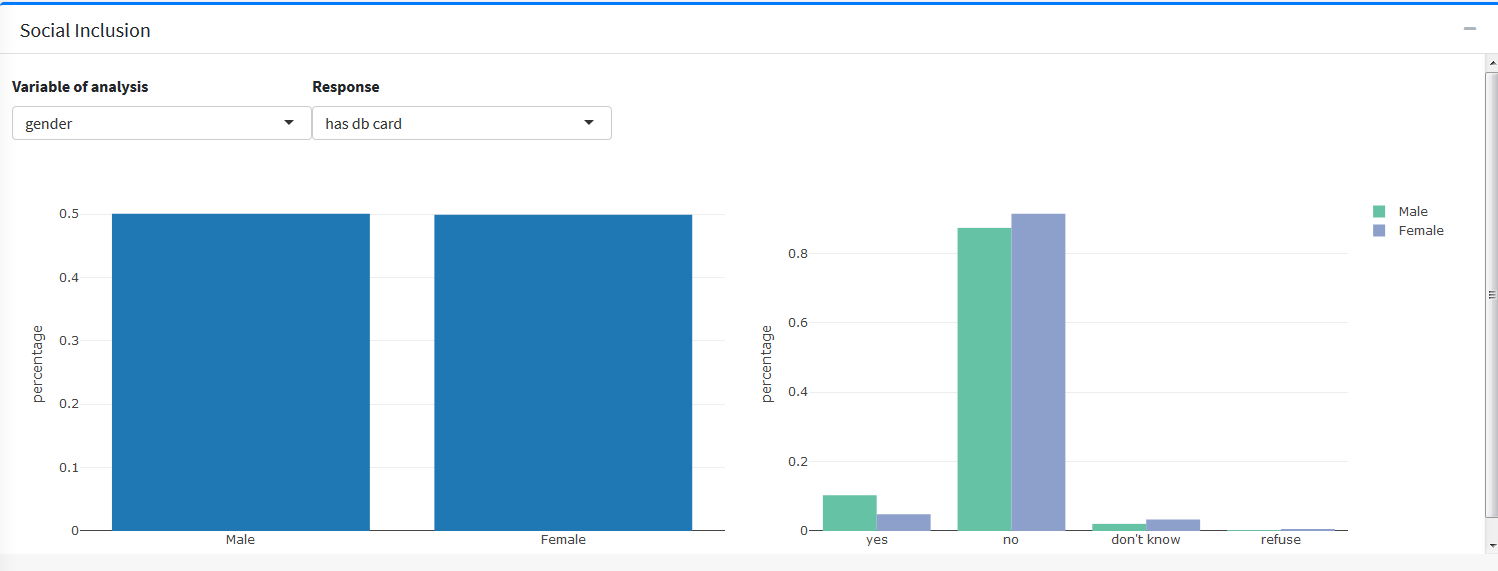
a) Social Inclusion

Cet onglet est sensible au sélecteur de la barre de gauche.

On trouve en haut une représentation des données issues du Global Financial Index.

Un premier sélecteur permet de choisir une variable sur laquelle on comparera les résultats par modalité.

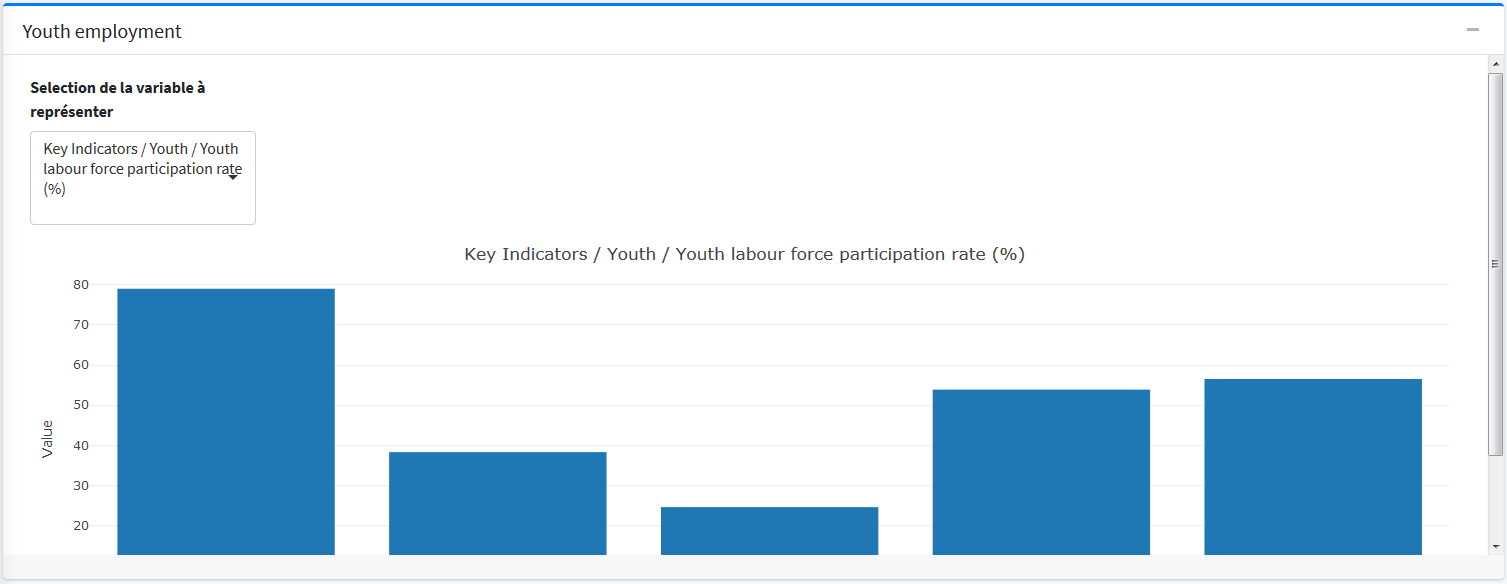
Le deuxième sélecteur permet de définir la variable numérique de réponse.



En dessous on trouve un graphique en barre permettant de visualiser 3 variables sur l'accès à l'emploi pour les jeunes

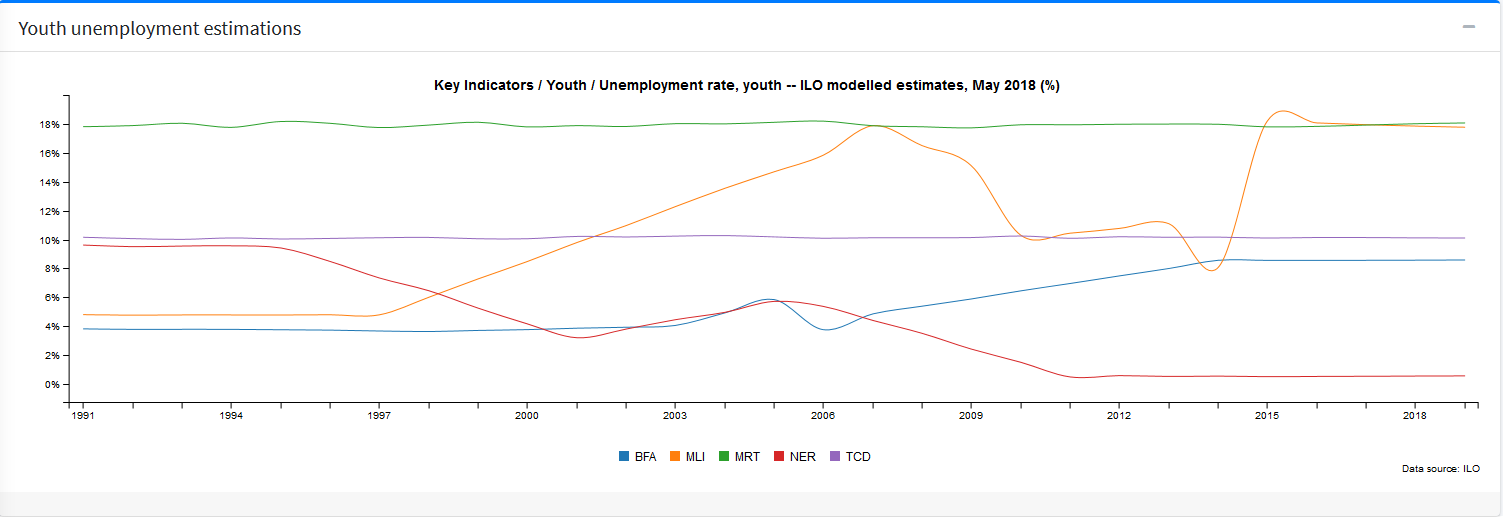
Tous les pays ne sont pas représentés car il peut y avoir des valeurs manquantes.

Ces données proviennent d'ILO.



On trouve en dessous des données estimée sur les tendances des taux de chômage depuis 1991.

Ces données proviennent également d'ILO.



b) Migration

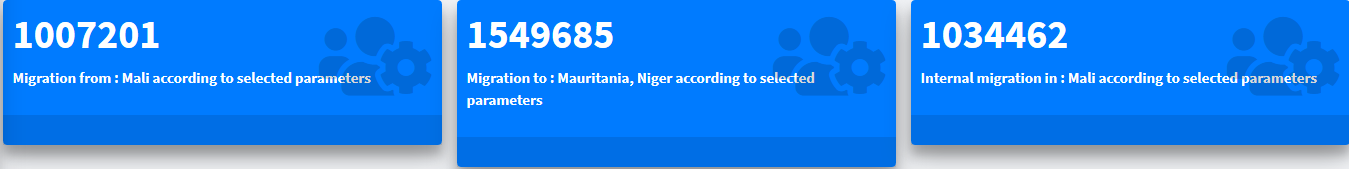
Cet onglet se base sur les données issues de l'UNHCR

On trouve en haut de cet onglet 3 KPI basés sur les contrôleurs sélectionnés en dessous.

- Le nombre de migrants partant

- Le nombre de migrants entrant

- Le nombre de migrants internes



Les contrôleurs permettent de sélectionner :

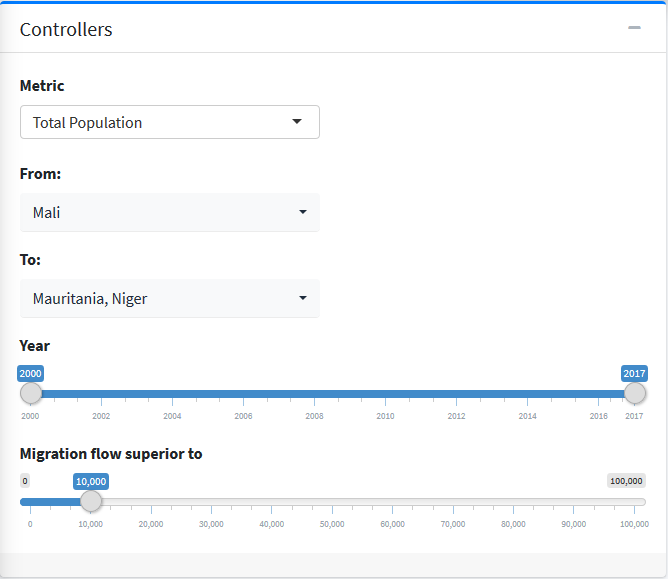
- La variable à représenter

- Les flux partants des pays sélectionnés par "From"

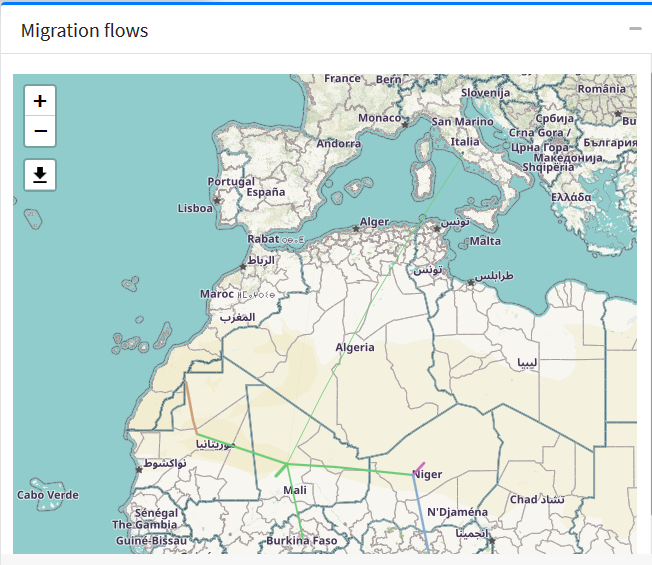
- Les flux entrants dans les pays sélectionnés par "To"

- La période temporelle

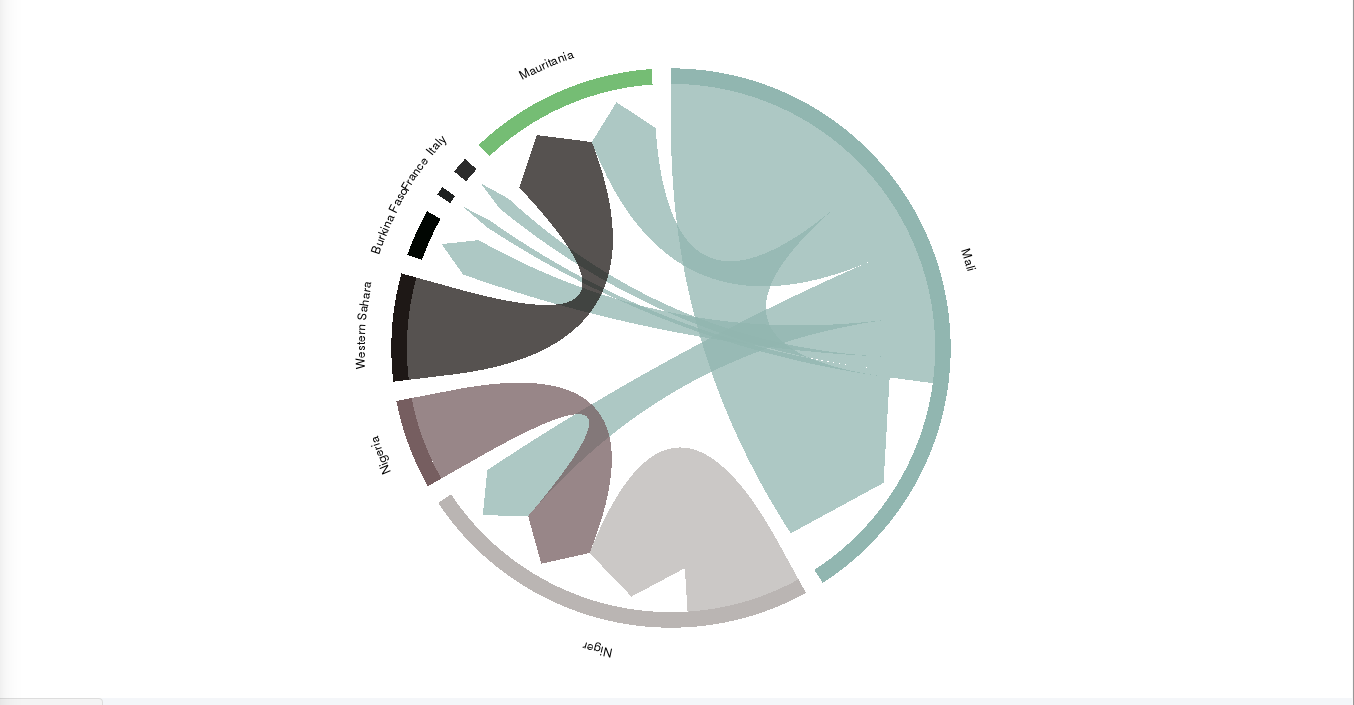
- L'application d'un filtre supprimant les petits flux



A la droite des contrôleurs, on trouve une carte interactive représentant les flux sélectionnés.



On trouve en dessous un Chord Diagram représentant les mêmes données.

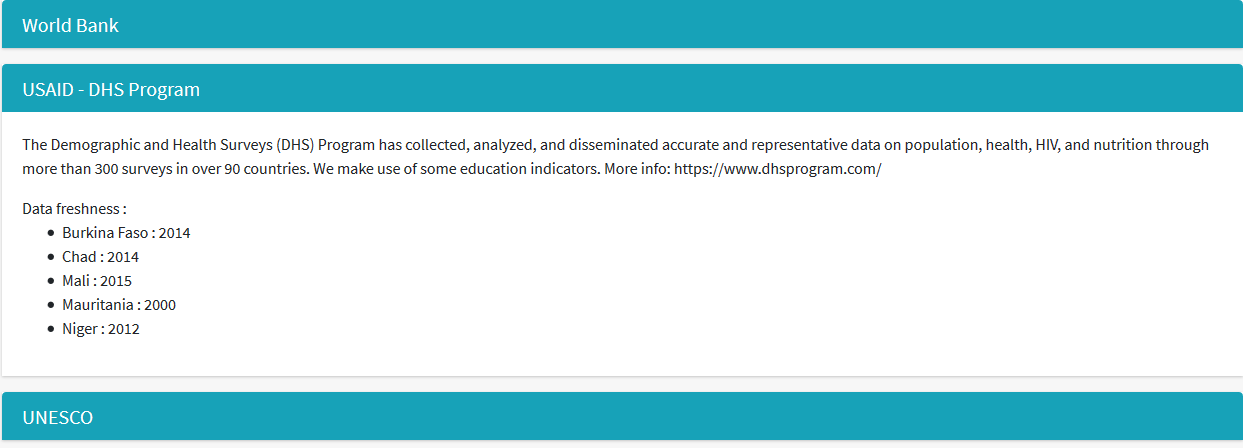


## D) Informations sur le projet :

a) Data Sources :

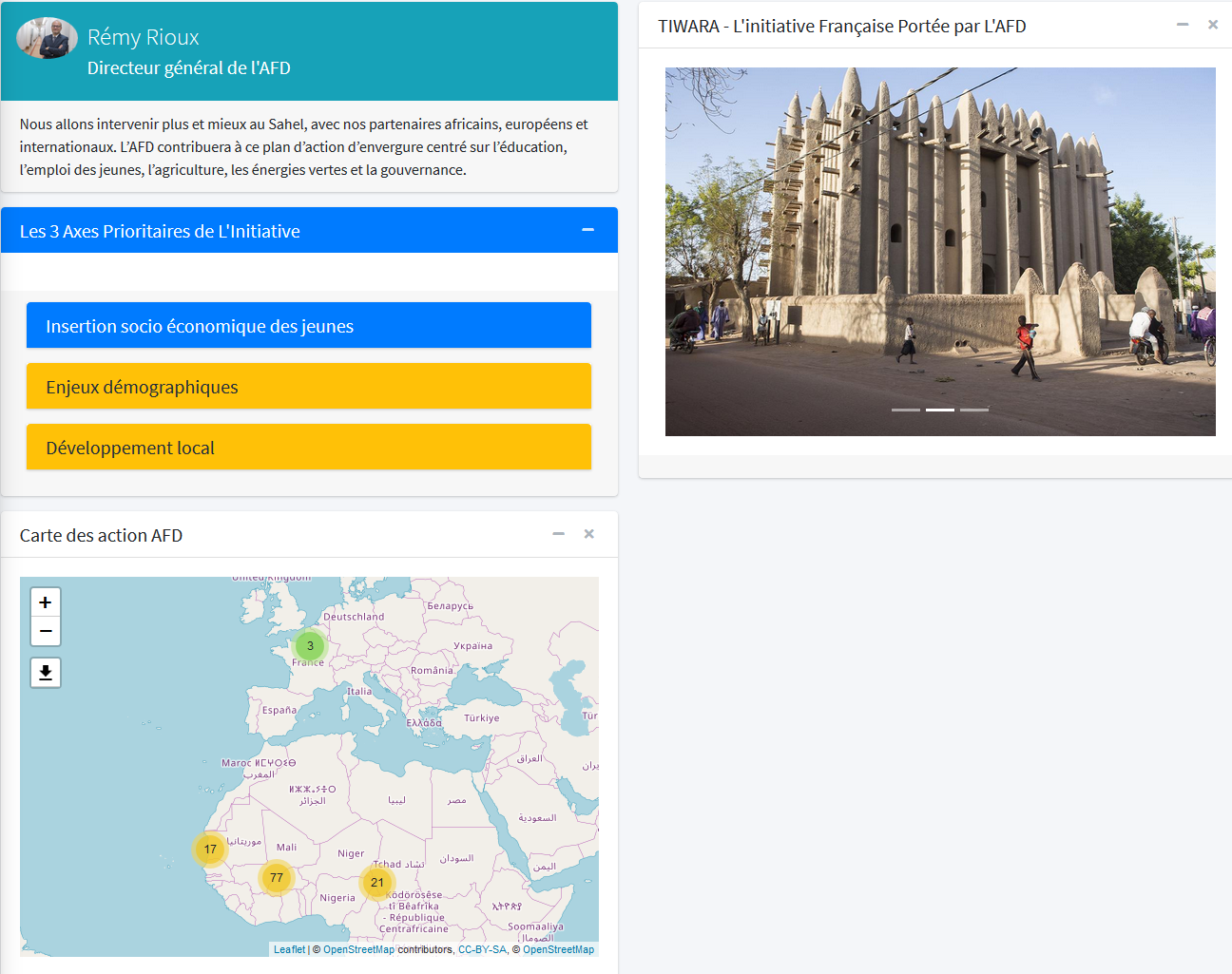
On trouve dans cet onglet des informations sur les différentes sources de données exploitées ainsi que des éléments sur les API permettant de les récupérer.

On trouve également l'information sur la fraîcheur des données (date de recueil).



b) A propos de Tiwara

Quelques informations sur l'initiative, les grands axes du projet ainsi qu'une cartographie interactive des projets de l'AFD basée sur les données ouvertes AFD.



# 3) La structure informatique du projet

Tiwara est développé en R. L'objectif étant de proposer un dashboard interactif, on utilise le package Shiny.

## A) La structure

Le projet est constitué d'un script principal appelé App.R contenant la majorité du code de l'application.

Dans le même répertoire que ce script, on trouve un fichier nommé cfg\_shinyapps.ini. Ce fichier contient les différents paramètres :

SAVE\_DIR : Emplacement du répertoire contenant les fichiers de données

[NOM\_FICHIERS] : Nom des fichiers de données

var\_dhs : Nom des variable du fichier DHS exploitées

SAHEL\_SHAPEFILE : Fichier contenant les polygones des régions du Sahel

[DBNOMICS\_EMPLOYMENT] : paramètres permettant de se connecter à l'API Dbnomics pour récupérer les données ILO

ISO3\_CN : Liste des pays du Sahel au format ISO3

ISO3\_CLUST : Liste des pays utilisés pour le clustering, format ISO3

COUNTRY\_NAME : Nom des pays du Sahel en anglais

COUNTRY\_NAME\_FR : Nom des pays du Sahel en français et en majuscule

Dans le futur, il faudra héberger le projet sur un git.

Le développement s'est effectué sur une instance RServer basée sur un RedHat Enterprise.

Le projet à été déployé sur Shinyapps.io ainsi que sur l'instance RConnect de l'AFD.

## B) Précisions sur le code

Liste des packages utilisés :

- shiny

- fontawesome

- shinyWidgets

- bs4Dash

- plotly

- r2d3

- r2d3maps

- rnaturalearth

- magick

- leaflet

- leaflet.extras

- lubridate

- DT

- recipes

- highcharter

- tidyverse

- billboarder

- echarts4r

- countrycode

- shinydashboard

- flexdashboard

- sf

- shinythemes

- htmlwidgets

- janitor

- configr

- httr

- jsonlite

- shinyWidgets

- corrplot

- tsne

- NbClust

- readxl

- readr

- geosphere

- qualpalr

- circlize

- rdbnomics

Le type d'application shiny utilisée provient du package bs4Dash.

Le chargement des données se fait dans l'en-tête de App.R (Au dessus de l'UI).

Le clustering est réalisé à l'aide du package NbClust qui calcule différents critères de qualité et les agrège pour déterminer le meilleur nombre de classes.

Les cartographies ont été réalisées avec les packages suivant :

- r2d3maps

- leaflet

Les autres packages graphiques utilisés :

- plotly

- billboarder

- highcharter

- circlize

Un profil de performance a été réalisé avec le package profvis. Il révèle que les graphiques prenant le plus de temps à être générés sont les cartographies de r2d3maps.

## C) Les mises à jour des données

La mise à jour des données n'est pas évidente dans le cadre du projet Tiwara.

La principale cause est la diversité des sources utilisées pour l'application. Ces sources ne sont pas mises à jour au même moment, et la récupération ne se fait pas de la même manière.

De plus certaines sources demandent un compte pour y avoir accès (DHS program, Data Financial Index).

Avancement de l'automatisation de la récupération des données :

Mise à jour par API :

- AFD

- ILO

- DHS (en utilisant RDHS ainsi qu'un compte autorisé?)

Récupération par requête GET ou scraping:

- UNHCR

- ACLED

- INFORM

A Traiter :

- Data Financial Index

- WDI