

# Arborescence

DAN & BRIGGS

13 octobre 2018  
Cr   par : Philippe Lawson

# Arborescence

---

## Data Lake

### Arborescence

- /data/frwiki
- /data/frwiki/raw
- /data/frwiki/frwiki-20180501
- /data/frwiki/frwiki-20180501/master
- /data/frwiki/frwiki-20180501/master/history.avsc
- /data/frwiki/frwiki-20180501/master/pagelinks.avsc
- /data/frwiki/frwiki-20180501/master/full
- /data/frwiki/frwiki-20180501/master/test

### Détail

#### **/data/frwiki**

Répertoire racine des données relatives au Wikipédia en français.

#### **/data/frwiki/raw**

Répertoire contenant les données brutes extraites du Wikipédia en français.

#### **/data/frwiki/frwiki-20180501**

Répertoire racine des données sérialisées du Wikipédia en français correspondant à l'extraction du 01/05/2018.

#### **/data/frwiki/frwiki-20180501/master**

Répertoire contenant le master dataset ainsi que les schémas de sérialisation.

#### **Schémas de sérialisation**

- history.avsc : Schéma de sérialisation Avro de l'historique des révisions.
- pagelinks.avsc : Schéma de sérialisation Avro des liens entre pages.

#### **Datasets composant le master dataset**

- /data/frwiki/frwiki-20180501/master/full : Dataset contenant l'intégralité des données sérialisées de l'extraction du 01/05/2018.
- /data/frwiki/frwiki-20180501/master/test : Dataset contenant un sous-ensemble du dataset full. Il est utilisé pour la mise au point des programmes d'analyse.

### Noms des fichiers

- revisions.<nnnn>.avro : Fichiers contenant l'historique des révisions.
- pagelinks.<nnnn>.avro : Fichiers contenant les liens entre pages.

<nnnn> est le numéro du fichier. Pour l'extraction du 01/05/2018, l'historique des révisions a été sérialisé en 5965 fichiers ; et les liens entre pages ont été sérialisés en 52 fichiers.