cikisi

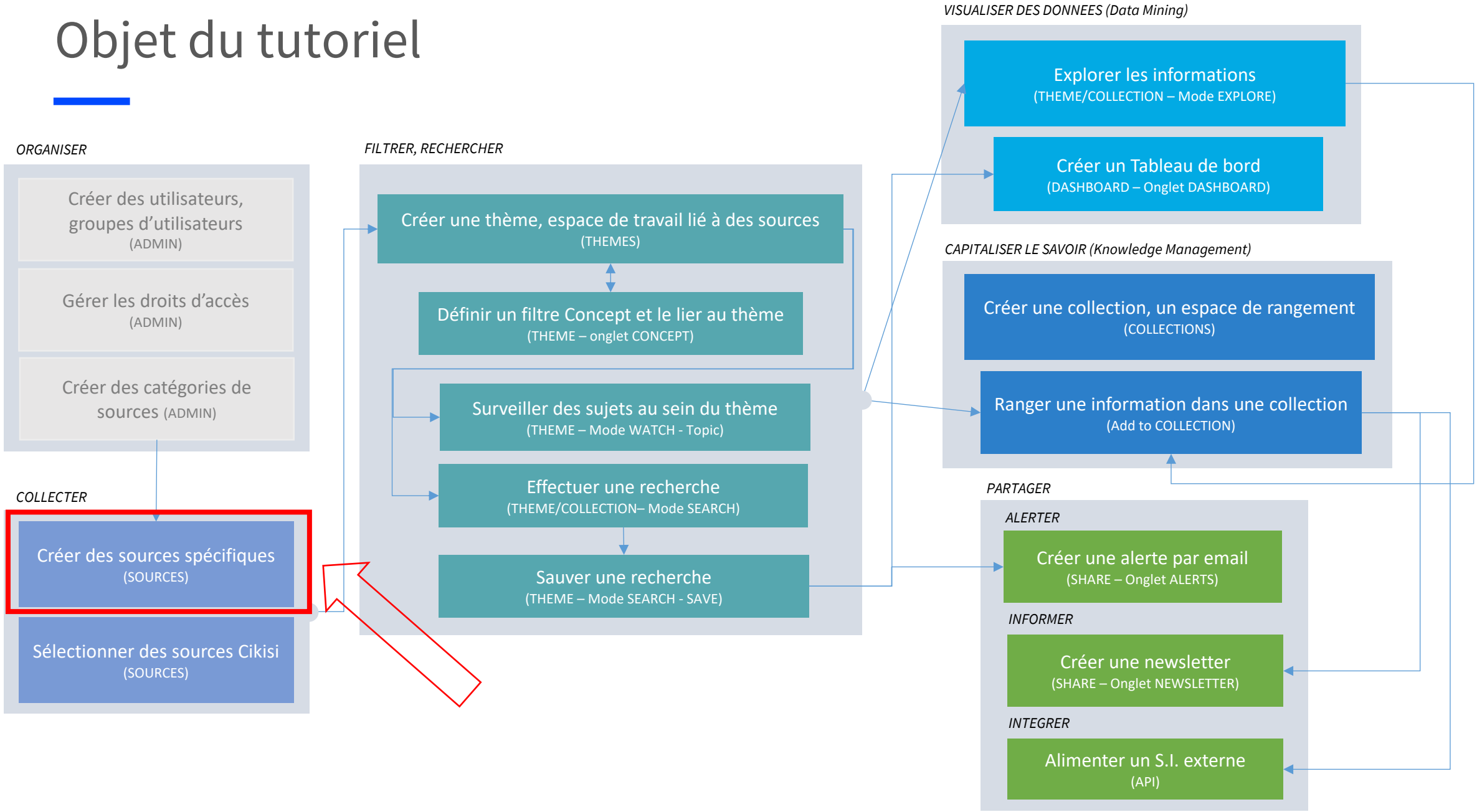
Search, Watch, Explore

Sources – Robot de collecte partant d’un flux (geo)RSS

V1.0

March 2021

Objet du tutoriel



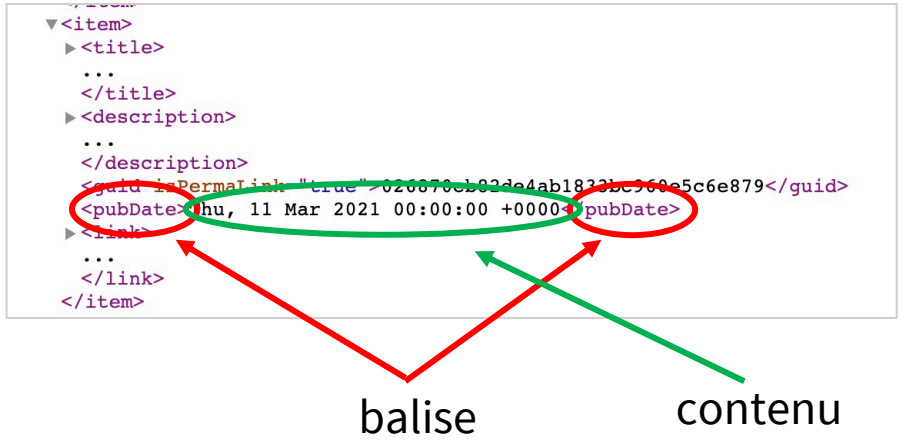
Plan

- **Définition** d’un flux (geo)RSS
- **Cas d’emploi**
- Vérification de la **validité**
- **Avantages**
- **Limitations**
- **Principes techniques**
- **Fréquence** de collecte
- **Trouver** le(s) flux RSS d’un site web

Définition d'un flux (geo)RSS

- Un flux RSS est un **fichier XML**, composé de balises encadrant du contenu.
- Les **balises** indiquent le type de contenu :
  - Titre
  - Description
  - Auteur
  - Date de publication
  - Lien
  - Lien de l'image associée
- Toutes les balises ne sont pas toujours présentes
- Les flux **geoRSS** sont ceux possédant également une balise reprenant des **coordonnées géographiques** (lat/long)
- Les symboles du flux RSS les plus fréquents sont les suivants :

XML = contenu structuré



Cas d'emploi

- Le type de source « Flux geo(RSS) » **est utilisé lorsque**:
  - Un ou des **flux RSS sont disponibles** sur le site à mettre en surveillance
  - Les **flux RSS sont bien alimentés**, de façon régulière et exhaustive par le site
  - Les articles listés au sein du flux RSS couvrent une période au minimum égale à la fréquence de collecte
    - Ne pas utiliser un robot de collecte partant d'un flux RSS avec une fréquence de 12 heures si les articles listés au sein du flux RSS sont relatifs aux 2 dernières heures. Sans quoi, la collecte sera partielle (1/6ème)

- Le type de source « Flux geo(RSS) » **ne doit pas être utilisé lorsque** le flux RSS provient d'une alerte **Google**, d'un site **Reddit** ou **Blogspot** (utiliser le type de source Google Alert, Reddit ou Blogspot)

Vérification de la validité d'un flux (geo)RSS

Un flux RSS est valide si :

- Il contient plusieurs articles
- La date de publication des articles est récente ou correspond à la dernière publication du site

Flux RSS pas valide

navy.gov.au/rss.xml

```
<?xml version="1.0" encoding="utf-8"?>
<rss version="2.0" xml:base="https://www.navy.gov.au"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<channel>
 <title>Royal Australian Navy</title>
 <link>https://www.navy.gov.au</link>
 <description></description>
 <language>en</language>
</channel>
</rss>
```

foodingredientsfirst.com/Content/RSS/fifnews.xml

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
<rss xmlns:atom="http://www.w3.org/2005/Atom" version="2.0">
<channel>
<title>FoodIngredients First</title>
<description>FoodIngredients First News</description>
<link>https://www.foodingredientsfirst.com</link>
<atom:link href="https://www.foodingredientsfirst.com" rel="self" type="application/rss+xml"/>
<item>
<title>
<![CDATA[ Natural nutrients: Allmicroalgae demonstrates "future-proof potential" of microalgae ]]>
</title>
<description>
<![CDATA[ <p>11 Mar 2021 --- The COVID-19 pandemic has changed the paradigm of food distribution, forcing quick adjustments to global supply chains, without setting aside environmental concerns. </p><p>In line with these industry challenges, Allmicroalgae, a Portugese biotech company, is addressing the full potential of algae as a natural versatile ingredient at a time when quality and transparency is crucial. </p> ]]>
</description>
<guid isPermaLink="true">b9290d46c7c44cdcbe566be22a9c7d44</guid>
<pubDate>Thu, 11 Mar 2021 00:00:00 +0000</pubDate>
<link>
<![CDATA[ https://www.foodingredientsfirst.com/news/from-farm-to-fork-allmicroalgae-demonstrates-future-proof-potential-of-microalgae.html ]]>
</link>
</item>
<item>
<title>
<![CDATA[ Cell-based divide: Cultivated seafood proponents push back at "lab-grown" labeling term ]]>
</title>
<description>
<![CDATA[ <p>11 Mar 2021 --- The cultivated seafood industry is in its infancy, and its products are still poised for market entry. As such, industry's cell-based proponents are cautioning against prematurely imposing regulations that would limit cultivated seafood labeling to terms that may distance consumers, such as "lab-grown." </p><p>Non-profit Good Food Institute (GFI) is urging that cultivated seafood producers should be allowed to use terms like "fillet" and "steak" in labeling products that have the structural and textural characteristics associated with those terms. </p><p>Meanwhile, aquaculture player BlueNalu is championing the term "cell-cultured" to replace other "disparaging and misleading" labels.</p><p><strong><em>FoodIngredientsFirst </em></strong>has reached out to both organizations for comment.</p> ]]>
</description>
<guid isPermaLink="true">f2319030c5de461f93b599fe34aad305</guid>
<pubDate>Thu, 11 Mar 2021 00:00:00 +0000</pubDate>
<link>
<![CDATA[ https://www.foodingredientsfirst.com/news/cell-based-divide-cultivated-seafood-proponents-push-back-at-lab-grown-labeling-term.html ]]>
</link>
</item>
<item>
<title>
<![CDATA[ Bühler's digitally integrated mill pegged as "major leap" to meet quality rice demand ]]>
</title>
<description>
<![CDATA[ <p>11 Mar 2021 --- The first rollout of Bühler's "smart" rice mills is entering the final stages with Swiss rice miller Riseria Taverne among the first operators to successfully implement the system. <br><br>"First results indicate significant cost savings, reduction in energy consumption and increased quality output," says Stefania Dolci, head of quality assurance at Riseria Taverne, Switzerland's largest rice mill.</p> ]]>
</description>
<guid isPermaLink="true">026870cb82de4ab1833bc960e8c6e879</guid>
<pubDate>Thu, 11 Mar 2021 00:00:00 +0000</pubDate>
<link>
<![CDATA[ https://www.foodingredientsfirst.com/news/buhlers-digitally-integrated-mill-pegged-as-major-leap-to-meet-quality-rice-demand.html ]]>
</link>
</item>
<item>
<title>
<![CDATA[ Salt, fat and MSG reformulation comes to the fore in savory snacking, says Intersnack exec ]]>
</title>
<description>
<![CDATA[ <p>11 Mar 2021 --- During the COVID-19 pandemic, the number of snacking "moments" throughout the day has increased as consumers grapple with eating occasions while working and spending more time at home. Taste continues to be a key driver for savory snacks, but today consumers are also looking for less salt, no artificial flavors and more nutritional value from this category.</p> ]]>
</description>
<guid isPermaLink="true">0bddba931117410ba36a0b51eeaf1c18</guid>
<pubDate>Thu, 11 Mar 2021 00:00:00 +0000</pubDate>
<link>
<![CDATA[ https://www.foodingredientsfirst.com/news/salt-fat-and-msg-reformulation-comes-to-the-fore-in-savory-snacking-says-intersnack-exec.html ]]>
</link>
</item>
<item>
<title>
<![CDATA[ Vegan and flexitarian consumers "dig deeper" when seeking plant-based products, say suppliers ]]>
</title>
<description>
<![CDATA[ <p dir="ltr">11 Mar 2021 --- Global consumers are "digging deeper beyond vegan claims" in their search for plant-based produce, which has led to increased pressure for manufacturers to innovate with trusted ingredients. Still, taste and texture come out on top when it comes to meat analogs and dairy alternatives.</p>
]]>
```

Avantages

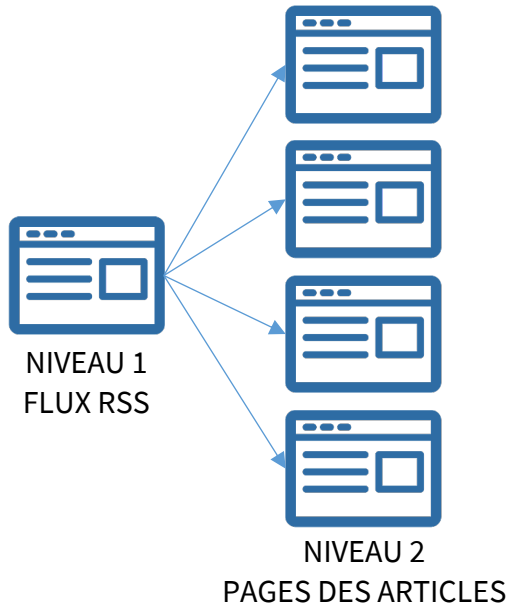
- **Configuration extrêmement simple**
  - Seule l'adresse du flux RSS est nécessaire
- En cas de changement de structure de site, **pas besoin de maintenance** (contrairement aux « scraping bots ») tant que le flux RSS existe toujours
- Permet de **collecter l'entièreté de l'article** et pas uniquement la description disponible dans le flux RSS
- **Fréquence de collecte élevée**
  - Jusqu'à 60 min
  - = en moyenne un article est collecté 30 min après sa publication

Limitations

- Parfois **le(s) flux RSS du site ne liste(nt) pas toutes les nouvelles pages créées**
- Si le **site web n'utilise pas des standards HTML** (de plus en plus rare), risque de ne pas reconnaître le contenu principal de la page et de limiter la collecte à la description
- Dans de rares cas, la **date de publication** de l'article n'est pas présente ou pas déclarée correctement dans le flux RSS et sera donc manquante.
- L'**adresse du flux RSS peut changer** au bout de quelques années il faut donc la mettre à jour (suivre le statut des sources)

Principes techniques

- Le robot Cikisi **extrait du contenu sur 2 niveaux:**
  - Le contenu présent au sein du **fichier XML** (flux RSS)
    - Contenu structuré
  - Le contenu présent sur **les liens (URLs) listés** au sein du flux RSS
    - Extraction automatique du contenu en interprétant les balises HTML



- **Pas d'écrasement**
  - Cikisi conserve la version la plus ancienne des pages (URL)
  - Une page déjà visitée ayant mené à la création d'un article n'est pas mise à jour
    - A noter qu'à l'exception des modifications mineures du contenu d'un article (orthographe, faute de frappe), les autres mises à jour font souvent l'objet d'une modification de l'URL et seront donc collectées par Cikisi.

Fréquence de collecte

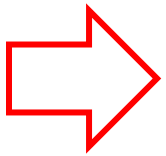
- Choisissez la fréquence de collecte en adéquation avec vos besoins en veille mais aussi en fonction de la quantité d'articles publiés par heure/jour par le site. En effet, le fichier XML contient uniquement les 10, 20 ou 50 (rarement) derniers articles publiés par le site. Visiter ce fichier pas assez fréquemment pourrait mener à rater des nouveaux articles.

- La fréquence de 60 minutes est idéale

Trouver le flux RSS d'un site

- **Lorsque le site déclare son flux RSS**
  - Chercher « RSS » au sein de la page à l'aide de CTRL-F
  - Identifier visuellement l'icône RSS quelque part sur la page
  - Regarder dans le pied de page sous le nom « RSS », « ATOM » ou « abonnez-vous »
  - Rendez-vous dans le section « news » / « actualités » où se situe fréquemment le flux

- **Lorsque le site possède un flux RSS mais ne le déclare pas**
  - Ajouter après le nom de domaine, un des chemins suivants :

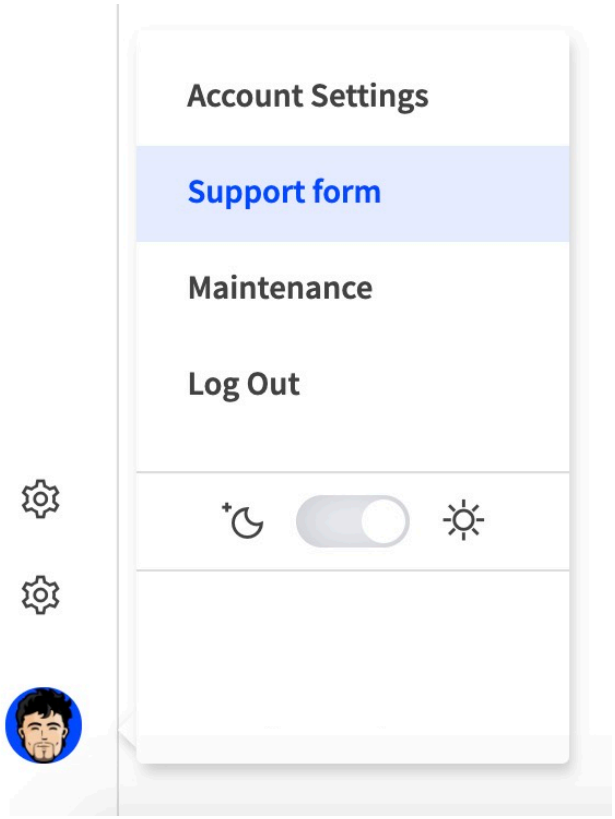


- **/feed**
- **/rss**
- **/rss.xml**
- /blog/feed
- /fr/feed
- news/rss
- news/feed
- rss.aspx
- feeds/posts/default
- rss.php
- /xml/syndication.rss
- /feed/rss
- fr/rss.xml
- /rss.jsp
- /rss/news
- /en/feed
- /rss/home
- /fr/rss
- /rss/2.0
- /category/news/feed
- en/rss.xml
- /index.rss
- /rss/feed
- /en/rss
- /feeds/news
- /rss/news.xml
- /blog/rss
- /rss.html
- blog/rss.xml
- /comments/feed
- /rss-feed
- /rss/all
- /news/rss.xml
- /rss/articles
- /feeds/all
- /feeds/rss
- /rss.business.xml
- /rss/news-releases.XML
- /feeds/rss.php
- …

Questions?

Utilisez notre formulaire de support!

https://wmt.cikisi.com/support



ATTENTION

Pour une compatibilité parfaite et une meilleure expérience utilisateur, Cikisi recommande l'emploi d'une version récente du navigateur Google Chrome