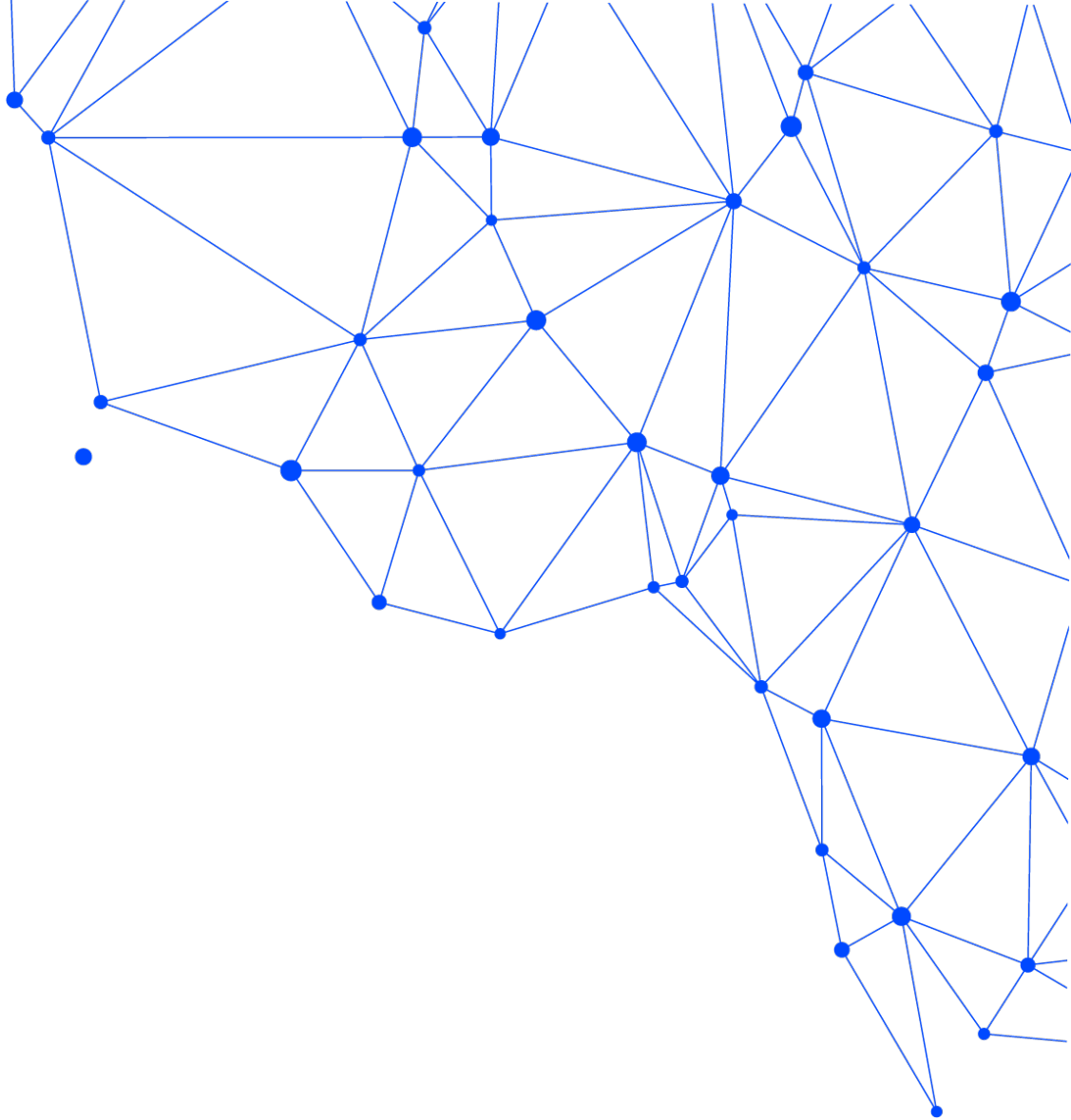


Sources - Website Crawler

V1.0

March 2021

Website Crawler – cas d'emploi

- Le type de source « website crawler » **est utilisé lorsque**:
  - Il n'est **pas possible de suivre les nouvelles publications** d'un site internet au travers d'un flux RSS, d'une ou plusieurs pages spécifiques pouvant être mise(s) en surveillance à l'aide d'un « scraping bot », d'un compte Twitter ou autre lié à ce site
  - On souhaite collecter des **publications plus anciennes** (non reprises au sein du flux RSS ou d'une page spécifique)
  - La **structure des pages** du site **change fréquemment** ou n'est pas uniforme, exigeant beaucoup de travail de configuration et de maintenance des sources de type « scraping bot »

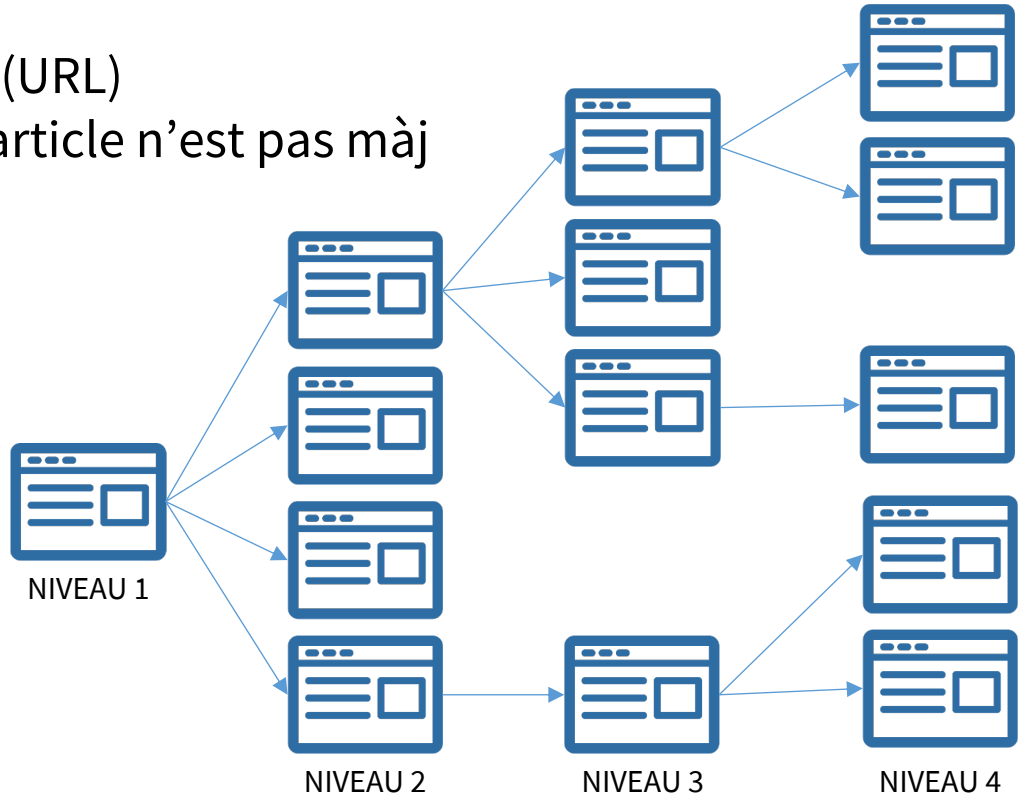
Website Crawler – avantages et désavantages

- **Avantages**
  - **Configuration extrêmement simple** – Cikisi détecte lui-même le contenu principal de la page à collecter
  - En cas de changement de structure de page, **pas besoin de maintenance** (contrairement aux « scraping bots »)
  - Permet une **collecte sur plusieurs niveaux** donc plus en profondeur dans le site (les autres types de sources travaillent sur 2 niveaux)

- **Désavantages**
  - Si le **site web n'utilise pas des standards HTML** (de plus en plus rare), risque de ne pas reconnaître le contenu principal du site et donc de ne rien collecter
    - Dans certains cas, la date de publication de l'article n'est pas déclarée correctement et sera donc manquante
  - **Fréquence de collecte moins élevée** que pour d'autres types de source (6 heures contre 60 minutes pour les flux RSS ou les « scraping bot »)
  - **Risque de détection** des robots Cikisi plus élevé et donc risque d'être bloqué, rendant la collecte impossible

Website Crawler – principes techniques

- Le crawler travaille sur **4 niveaux**, visitant les URLs trouvés au sein des pages
- Lors de sa première session de crawl, Cikisi va créer maximum 1,000 articles
- Lors des sessions de crawl ultérieures, Cikisi va créer maximum 200 articles
- **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 màj
- Crawl « éthique » et pas « massif »
  - Limitation du nombre de pages visitées/créées par session de crawl
  - Intervalle de temps entre deux pages visitées
  - Si deux clients Cikisi demandent à crawler la même source, celle-ci ne sera visitée qu'une seule fois



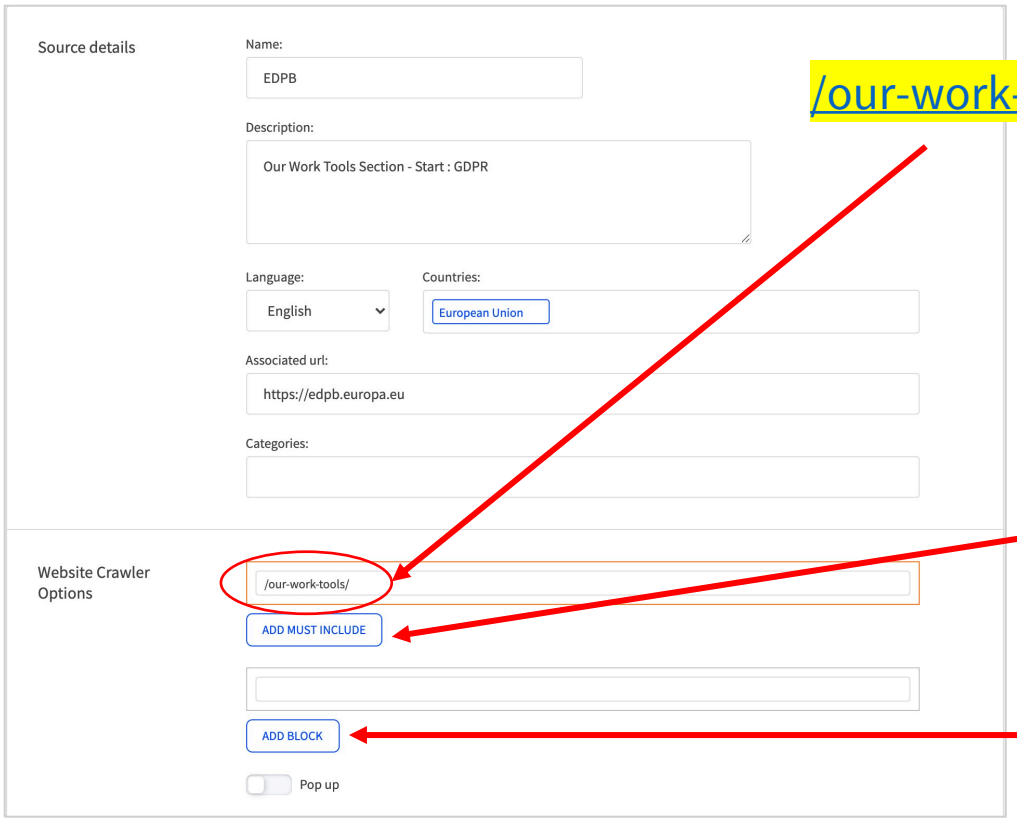
Website Crawler – définition du 1er niveau

- Le premier niveau est défini de la sorte :
  - Session de crawl impair : l'**URL de la page de démarrage** définie par l'utilisateur (« starting page ») lors d'une session de crawl sur deux
  - Session de crawl pair : l'**URL d'une autre page** ayant mené à la création d'un article, choisi **aléatoirement** (NEW)

- Cette technique permet de découvrir davantage de parties du site web, de ne pas emprunter des itinéraires de crawl toujours fort similaires

Website Crawler - Ciblage de certaines pages

Starting Page URL https://edpb.europa.eu/our-work-tools/general-guidance/gdpr-guidelines-recommendations-best-practices_en



/our-work-tools/

Il est possible de demander au crawler de Cikisi de ne collecter que les pages dont l'URL contient un **chemin spécifique**.

Les chemins fréquemment utilisés par les sites web contiennent :

- /news/ ou /actualites/
- /article/ ou /artikel/
- /post/
- /blog/

Cikisi permet de spécifier plusieurs chemins en cliquant sur le bouton « ADD MUST INCLUDE »

Exemple : collecter toutes les pages dont le chemin contient /news/france/ ou /news/monde/

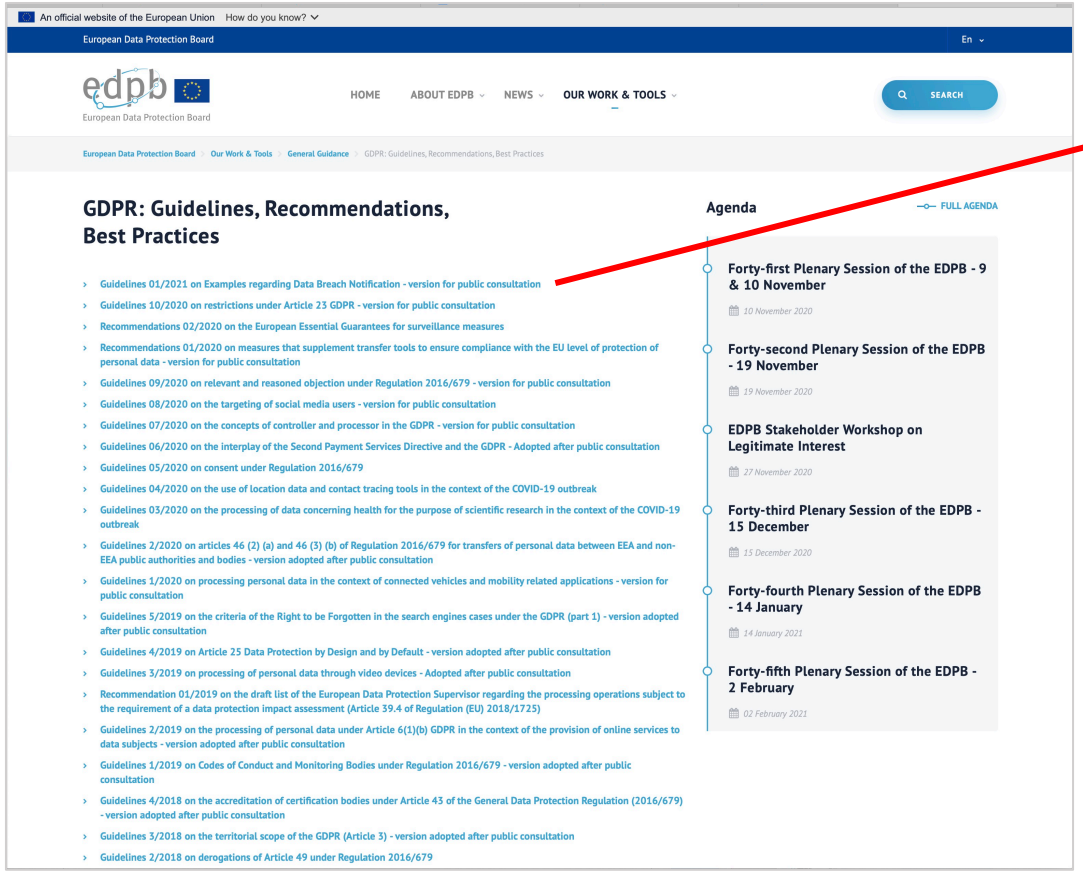
De même, il est possible de préciser les pages ne devant pas être collectées à l'aide du « ADD BLOCK »

Exemple : ne pas collecter les pages dont le chemin contient /meteo/

Website Crawler - Ciblage de certaines pages

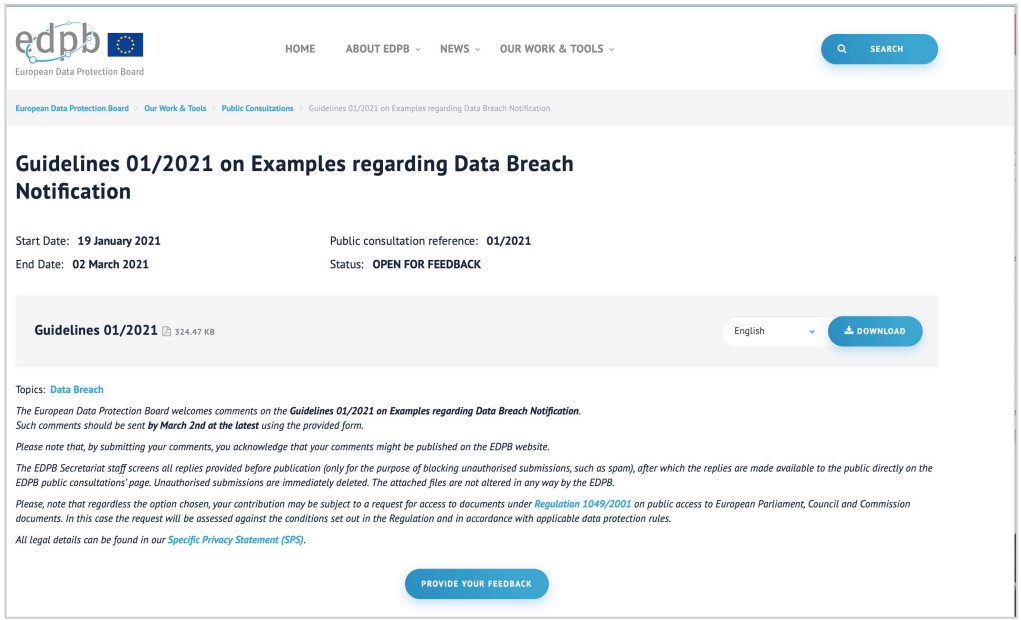
URL de départ (« starting page »)

https://edpb.europa.eu/our-work-tools/general-guidance/gdpr-guidelines-recommendations-best-practices_en

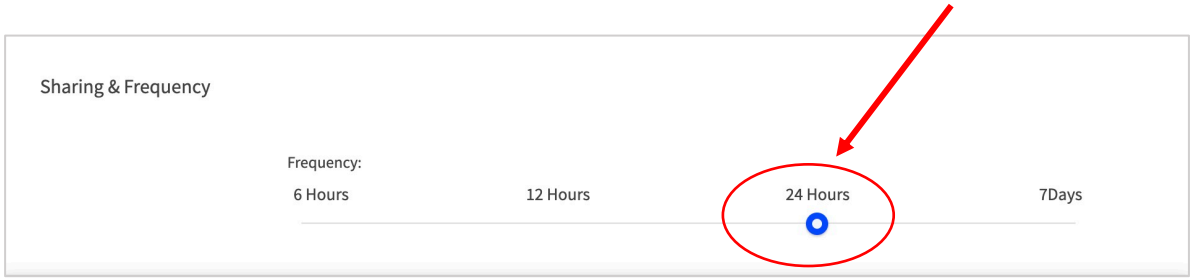


URL des pages à collecter doit toujours contenir un chemin donné :

https://edpb.europa.eu/our-work-tools/public-consultations-art-704/2021/guidelines-012021-examples-regarding-data-breach_en



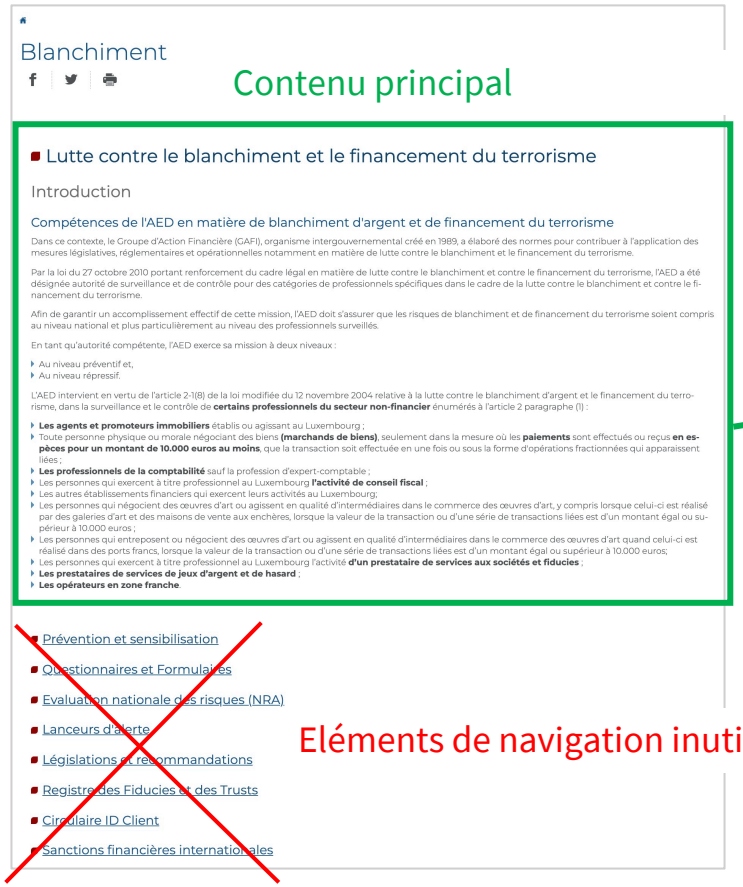
Website Crawler - Fréquence de collecte

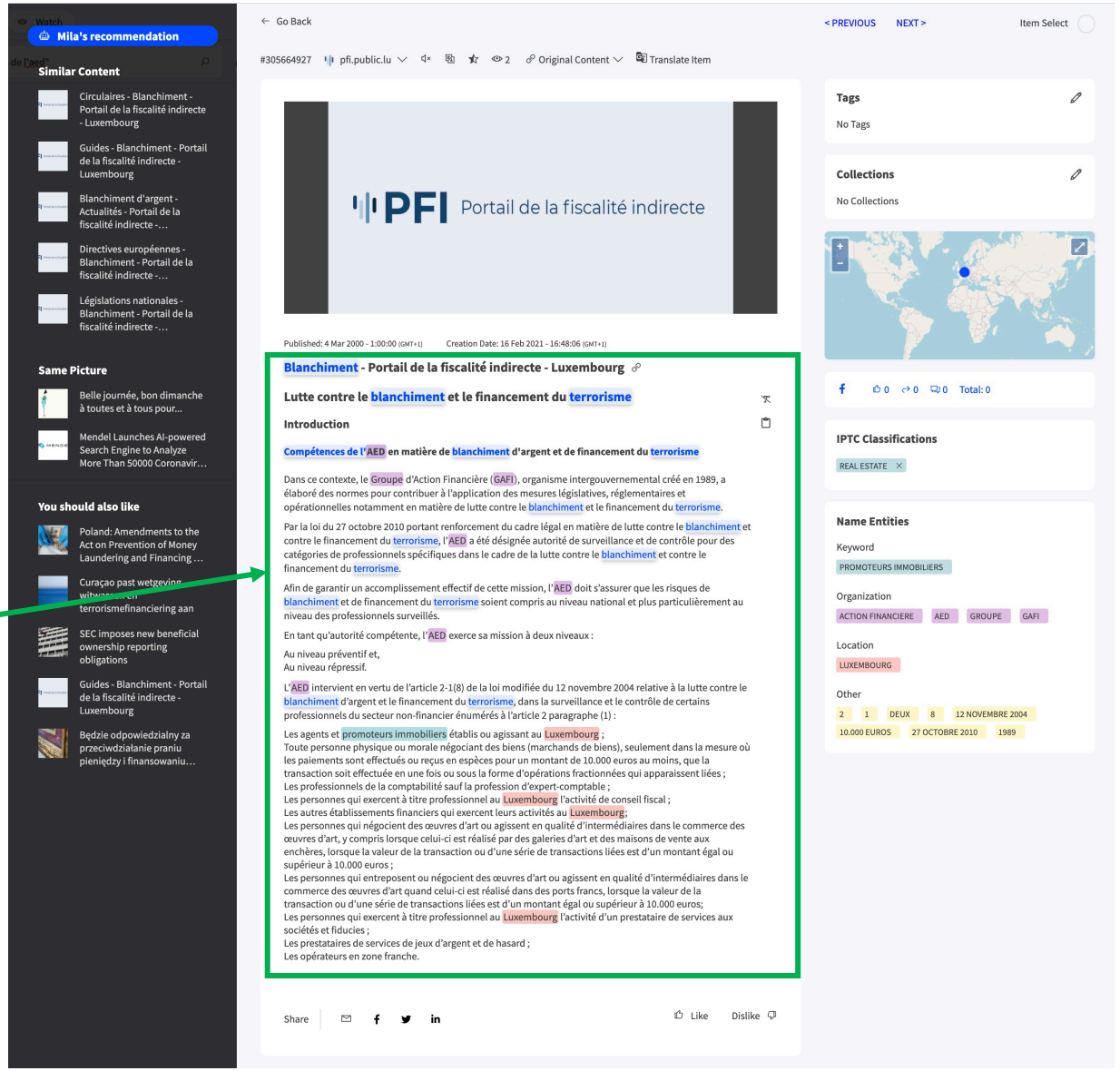


- Cikisi emploie des mesures d'**anonymisation** pour contourner 90 % des systèmes de détection de robots des sites web
- Pour éviter tout risque inutile de détection, il est recommandé d'**ajuster la fréquence à ses besoins** (inutile d'opter pour une fréquence de 6 heures si l'on procède à une veille hebdomadaire, voir même quotidienne)
- Il faut tenir compte du fait que le crawl à partir de la page de démarrage s'opère une session sur deux

Website Crawler – Contenu principal

Exemple de création automatique d'un article dans Cikisi à l'aide du « website crawler » lorsque les champs principaux sont bien renseignés

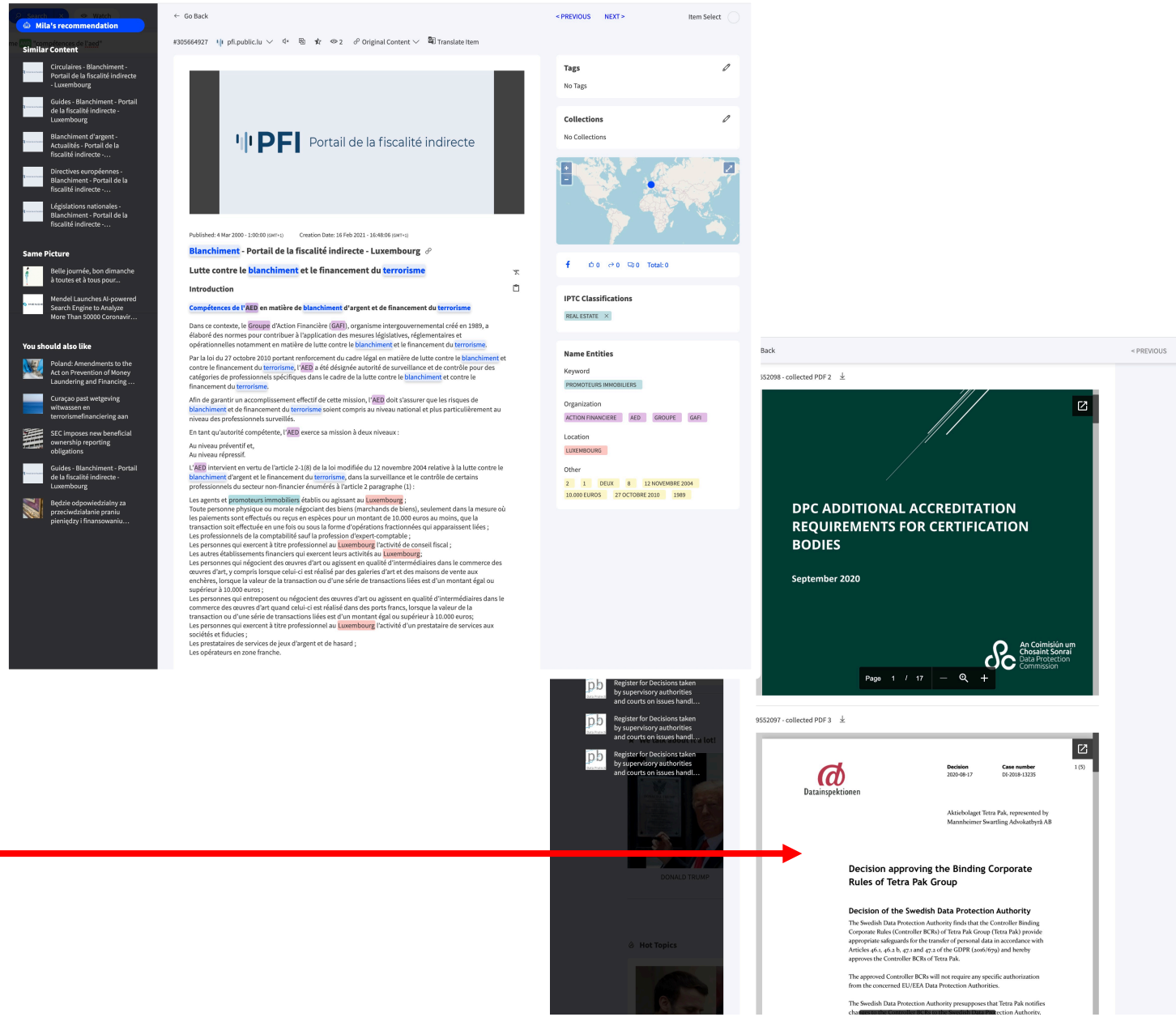




Website Crawler - PDF

Le website crawler détecte également les liens PDF insérés dans la page collectée.

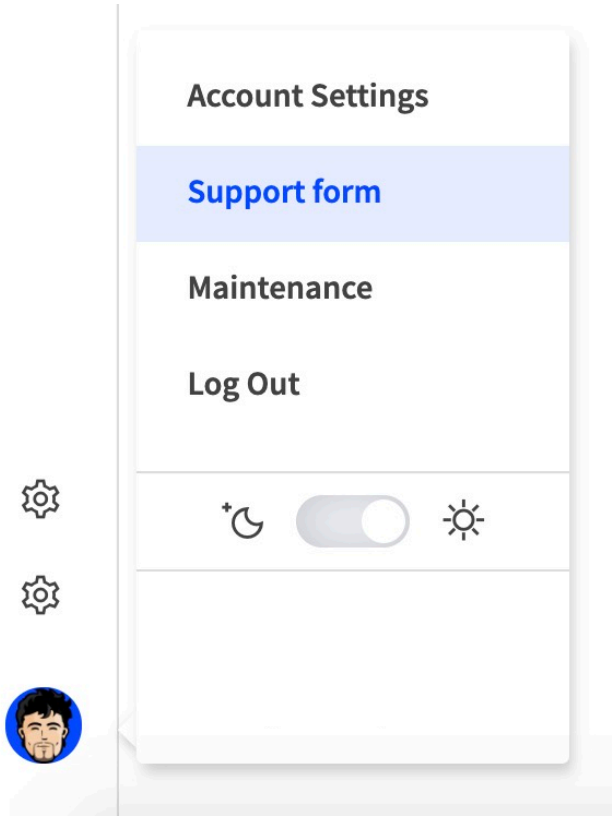
Cikisi permet ensuite à l'utilisateur de visionner les documents liés en dessous du contenu textuel de l'article (« view item »).



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