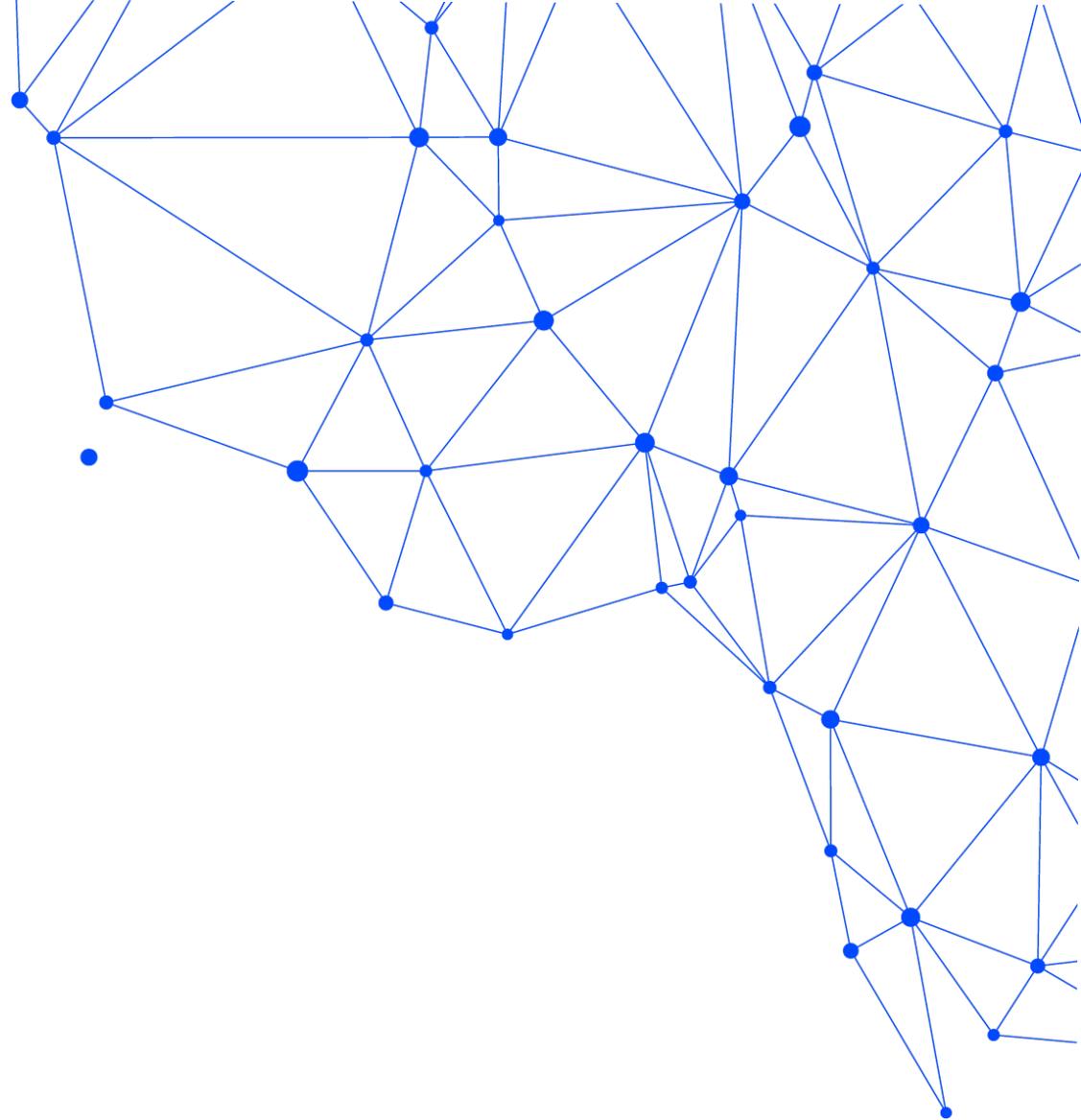


cikisi
Search, Watch, Explore

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

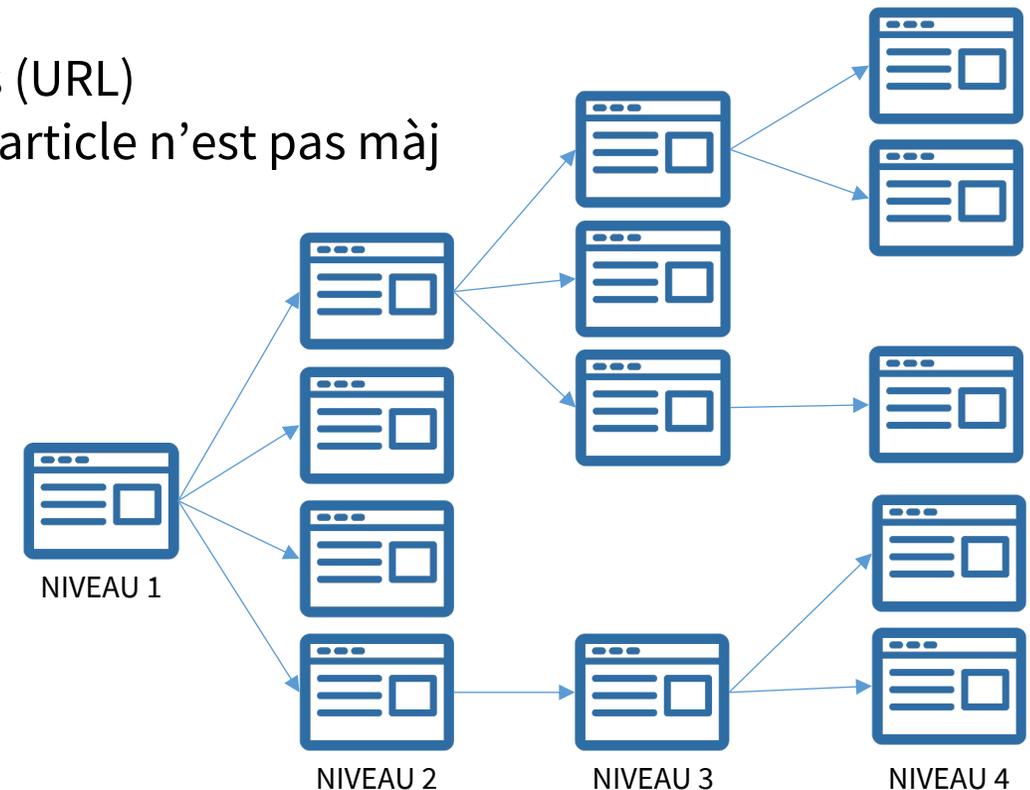
- **Configuration extrêmement simple** – Cikisi détecte lui-même le contenu principal de la page à collecter
- En cas de changement de structure de page, **pas besoin de maintenance** (contrairement aux « scraping bots »)
- Permet une **collecte sur plusieurs niveaux** donc plus en profondeur dans le 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 m à j
- 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 1^{er} 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

The screenshot shows the Cikisi website crawler configuration interface. It is divided into two main sections: 'Source details' and 'Website Crawler Options'.
In the 'Source details' section, there are fields for 'Name' (EDPB), 'Description' (Our Work Tools Section - Start : GDPR), 'Language' (English), 'Countries' (European Union), 'Associated url' (https://edpb.europa.eu), and 'Categories'.
In the 'Website Crawler Options' section, there is a text input field containing '/our-work-tools/'. This field is circled in red. Below it is a blue button labeled 'ADD MUST INCLUDE'. Further down, there is another empty text input field and a blue button labeled 'ADD BLOCK'. At the bottom, there is a 'Pop up' checkbox which is currently unchecked. A yellow highlight is placed over the text '/our-work-tools/' in the input field. Red arrows point from the text in the input field to the 'ADD MUST INCLUDE' button and from the 'ADD BLOCK' button to the right side of the slide.

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

The screenshot shows the EDPB website with the following structure:

- Header: European Data Protection Board, HOME, ABOUT EDPB, NEWS, OUR WORK & TOOLS, SEARCH.
- Breadcrumbs: European Data Protection Board > Our Work & Tools > General Guidance > GDPR: Guidelines, Recommendations, Best Practices.
- Main Content: **GDPR: Guidelines, Recommendations, Best Practices**. A list of links including "Guidelines 01/2021 on Examples regarding Data Breach Notification - version for public consultation".
- Agenda: A vertical timeline of events including "Forty-first Plenary Session of the EDPB - 9 & 10 November", "Forty-second Plenary Session of the EDPB - 19 November", "EDPB Stakeholder Workshop on Legitimate Interest", "Forty-third Plenary Session of the EDPB - 15 December", "Forty-fourth Plenary Session of the EDPB - 14 January", and "Forty-fifth Plenary Session of the EDPB - 2 February".

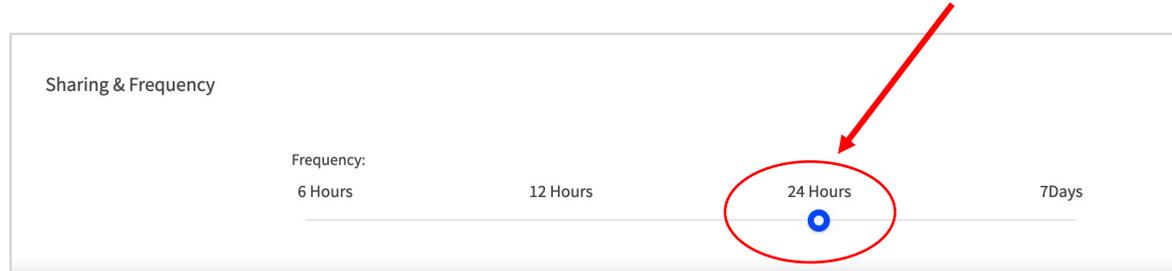
URL des pages à collecter doit toujours contenir un chemin donné :

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

The screenshot shows the EDPB website with the following structure:

- Header: European Data Protection Board, HOME, ABOUT EDPB, NEWS, OUR WORK & TOOLS, SEARCH.
- Breadcrumbs: European Data Protection Board > Our Work & Tools > Public Consultations > Guidelines 01/2021 on Examples regarding Data Breach Notification.
- Main Content: **Guidelines 01/2021 on Examples regarding Data Breach Notification**. Start Date: 19 January 2021, End Date: 02 March 2021, Public consultation reference: 01/2021, Status: OPEN FOR FEEDBACK.
- Download: Guidelines 01/2021 (324.47 KB), English, DOWNLOAD.
- Topics: Data Breach.
- Text: "The European Data Protection Board welcomes comments on the Guidelines 01/2021 on Examples regarding Data Breach Notification. Such comments should be sent by March 2nd at the latest using the provided form. Please note that, by submitting your comments, you acknowledge that your comments might be published on the EDPB website. The EDPB Secretariat staff screens all replies provided before publication (only for the purpose of blocking unauthorised submissions, such as spam), after which the replies are made available to the public directly on the EDPB public consultations' page. Unauthorised submissions are immediately deleted. The attached files are not altered in any way by the EDPB. Please, note that regardless the option chosen, your contribution may be subject to a request for access to documents under Regulation 1049/2001 on public access to European Parliament, Council and Commission documents. In this case the request will be assessed against the conditions set out in the Regulation and in accordance with applicable data protection rules. All legal details can be found in our Specific Privacy Statement (SPS)."
- Button: PROVIDE YOUR FEEDBACK.

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

Blanchiment

Contenu principal

- Lutte contre le blanchiment et le financement du terrorisme

Introduction

Compétences de l'AED en matière de blanchiment d'argent et de financement du terrorisme

Dans ce contexte, le Groupe d'Action Financière (GAFI), organisme intergouvernemental créé en 1989, a élaboré des normes pour contribuer à l'application des mesures législatives, réglementaires et opérationnelles notamment en matière de lutte contre le blanchiment et le financement du terrorisme.

Par la loi du 27 octobre 2010 portant renforcement du cadre légal en matière de lutte contre le blanchiment et contre le financement du terrorisme, l'AED a été désignée autorité de surveillance et de contrôle pour des catégories de professionnels spécifiques dans le cadre de la lutte contre le blanchiment et contre le financement du terrorisme.

Afin de garantir un accomplissement effectif de cette mission, l'AED doit s'assurer que les risques de blanchiment et de financement du terrorisme soient compris au niveau national et plus particulièrement au niveau des professionnels surveillés.

En tant qu'autorité compétente, l'AED exerce sa mission à deux niveaux :

- ▶ Au niveau préventif et,
- ▶ Au niveau répressif.

L'AED intervient en vertu de l'article 2-1(8) de la loi modifiée du 12 novembre 2004 relative à la lutte contre le blanchiment d'argent et le financement du terrorisme, dans la surveillance et le contrôle de **certaines professions du secteur non-financier** énumérés à l'article 2 paragraphe (1) :

- ▶ Les agents et promoteurs immobiliers établis ou agissant au Luxembourg ;
- ▶ Toute personne physique ou morale négociant des biens (marchands de biens), seulement dans la mesure où les paiements sont effectués ou reçus en espèces pour un montant de 10.000 euros au moins, que la transaction soit effectuée en une fois ou sous la forme d'opérations fractionnées qui apparaissent liées ;
- ▶ Les professionnels de la comptabilité sauf la profession d'expert-comptable ;
- ▶ Les personnes qui exercent à titre professionnel au Luxembourg l'activité de conseil fiscal ;
- ▶ Les autres établissements financiers qui exercent leurs activités au Luxembourg ;
- ▶ Les personnes qui négocient des œuvres d'art ou agissent en qualité d'intermédiaires dans le commerce des œuvres d'art, y compris lorsque celui-ci est réalisé par des galeries d'art et de maisons de vente aux enchères, lorsque la valeur de la transaction ou d'une série de transactions liées est d'un montant égal ou supérieur à 10.000 euros ;
- ▶ Les personnes qui entretiennent ou négocient des œuvres d'art ou agissent en qualité d'intermédiaires dans le commerce des œuvres d'art quand celui-ci est réalisé dans des ports francs, lorsque la valeur de la transaction ou d'une série de transactions liées est d'un montant égal ou supérieur à 10.000 euros ;
- ▶ Les personnes qui exercent à titre professionnel au Luxembourg l'activité d'un prestataire de services aux sociétés et fiduciaires ;
- ▶ Les prestataires de services de jeux d'argent et de hasard ;
- ▶ Les opérateurs en zone franche.

- Prévention et sensibilisation
- Questionnaires et Formulaires
- Evaluation nationale des risques (NRA)
- Lanceurs d'alerte
- Législations et recommandations
- Registres des Fiducies et des Trusts
- Circulaire ID Client
- Sanctions financières internationales

Éléments de navigation inutiles

Mila's recommendation

Similar Content

- Circulaires - Blanchiment - Portail de la fiscalité indirecte - Luxembourg
- Guides - Blanchiment - Portail de la fiscalité indirecte - Luxembourg
- Blanchiment d'argent - Actualités - Portail de la fiscalité indirecte - ...
- Directives européennes - Blanchiment - Portail de la fiscalité indirecte - ...
- Législations nationales - Blanchiment - Portail de la fiscalité indirecte - ...

Same Picture

- Belle journée, bon dimanche à toutes et à tous pour...
- Mendel Launches AI-powered Search Engine to Analyze More Than 50000 Coronavir...

You should also like

- Poland: Amendments to the Act on Prevention of Money Laundering and Financing ...
- Curacao past wetgeving wettigheidsaanpak tegen terrorismefinanciering aan
- SEC imposes new beneficial ownership reporting obligations
- Guides - Blanchiment - Portail de la fiscalité indirecte - Luxembourg
- Będzie odpowiedzialny za przeciwdziałanie praniu pieniędzy i finansowaniu...

Blanchiment - Portail de la fiscalité indirecte - Luxembourg

Lutte contre le blanchiment et le financement du terrorisme

Introduction

Compétences de l'AED en matière de blanchiment d'argent et de financement du terrorisme

Dans ce contexte, le Groupe d'Action Financière (GAFI), organisme intergouvernemental créé en 1989, a élaboré des normes pour contribuer à l'application des mesures législatives, réglementaires et opérationnelles notamment en matière de lutte contre le blanchiment et le financement du terrorisme.

Par la loi du 27 octobre 2010 portant renforcement du cadre légal en matière de lutte contre le blanchiment et contre le financement du terrorisme, l'AED a été désignée autorité de surveillance et de contrôle pour des catégories de professionnels spécifiques dans le cadre de la lutte contre le blanchiment et contre le financement du terrorisme.

Afin de garantir un accomplissement effectif de cette mission, l'AED doit s'assurer que les risques de blanchiment et de financement du terrorisme soient compris au niveau national et plus particulièrement au niveau des professionnels surveillés.

En tant qu'autorité compétente, l'AED exerce sa mission à deux niveaux :

Au niveau préventif et,

Au niveau répressif.

L'AED intervient en vertu de l'article 2-1(8) de la loi modifiée du 12 novembre 2004 relative à la lutte contre le blanchiment d'argent et le financement du terrorisme, dans la surveillance et le contrôle de certains professionnels du secteur non-financier énumérés à l'article 2 paragraphe (1) :

Les agents et promoteurs immobiliers établis ou agissant au Luxembourg ;

Toute personne physique ou morale négociant des biens (marchands de biens), seulement dans la mesure où les paiements sont effectués ou reçus en espèces pour un montant de 10.000 euros au moins, que la transaction soit effectuée en une fois ou sous la forme d'opérations fractionnées qui apparaissent liées ;

Les professionnels de la comptabilité sauf la profession d'expert-comptable ;

Les personnes qui exercent à titre professionnel au Luxembourg l'activité de conseil fiscal ;

Les autres établissements financiers qui exercent leurs activités au Luxembourg ;

Les personnes qui négocient des œuvres d'art ou agissent en qualité d'intermédiaires dans le commerce des œuvres d'art, y compris lorsque celui-ci est réalisé par des galeries d'art et des maisons de vente aux enchères, lorsque la valeur de la transaction ou d'une série de transactions liées est d'un montant égal ou supérieur à 10.000 euros ;

Les personnes qui entretiennent ou négocient des œuvres d'art ou agissent en qualité d'intermédiaires dans le commerce des œuvres d'art quand celui-ci est réalisé dans des ports francs, lorsque la valeur de la transaction ou d'une série de transactions liées est d'un montant égal ou supérieur à 10.000 euros ;

Les personnes qui exercent à titre professionnel au Luxembourg l'activité d'un prestataire de services aux sociétés et fiduciaires ;

Les prestataires de services de jeux d'argent et de hasard ;

Les opérateurs en zone franche.

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 »).

The screenshot displays a web browser interface with a search result for "Blanchiment - Portail de la fiscalité indirecte - Luxembourg". The main content area shows the title "Blanchiment - Portail de la fiscalité indirecte - Luxembourg" and a sub-heading "Lutte contre le blanchiment et le financement du terrorisme". Below this, there is an introduction section titled "Introductions de l'AED en matière de blanchiment d'argent et de financement du terrorisme". The text discusses the role of the AED (Action Financière) in combating money laundering and terrorism financing, mentioning the 2010 law and the 2004 law. It lists various categories of professionals and activities that are monitored, such as real estate agents, accountants, and art dealers.

On the left side, there is a sidebar with "Similar Content" and "Same Picture" sections. The "Similar Content" section lists several documents related to money laundering and terrorism financing, including "Circulaires - Blanchiment - Portail de la fiscalité indirecte - Luxembourg" and "Directives européennes - Blanchiment - Portail de la fiscalité indirecte - Luxembourg". The "Same Picture" section shows a list of images, including "Belle journée, bon dimanche à toutes et à tous pour..." and "Mandel Launches AI-powered Search Engine to Analyze More Than 50000 Coronavirus...".

On the right side, there is a "Tags" section with "No Tags", a "Collections" section with "No Collections", and a "Name Entities" section with a table of entities. The table has columns for "Keyword", "Organization", "Location", and "Other". The "Other" column contains a table with the following data:

2	1	DEUX	8	12 NOVEMBRE 2004
10.000 EUROS				27 OCTOBRE 2010
				1989

Below the main content, there is a list of related PDF documents. The first document is "DPC ADDITIONAL ACCREDITATION REQUIREMENTS FOR CERTIFICATION BODIES" by An Coimisiún um Chosaint Sonraí Data Protection Commission, dated September 2020. The second document is "Decision approving the Binding Corporate Rules of Tetra Pak Group" by the Swedish Data Protection Authority, dated 2020-09-17. The document text states: "The Swedish Data Protection Authority finds that the Controller Binding Corporate Rules (Controller BCRs) of Tetra Pak Group (Tetra Pak) provide appropriate safeguards for the transfer of personal data in accordance with Articles 46.1, 46.2 b, 47.1 and 47.2 of the GDPR (2016/679) and hereby approves the Controller BCRs of Tetra Pak. The approved Controller BCRs will not require any specific authorization from the concerned EU/EEA Data Protection Authorities. The Swedish Data Protection Authority presupposes that Tetra Pak notifies the concerned EU/EEA Data Protection Authorities."

Questions?

Utilisez notre formulaire de support!

<https://wmt.cikisi.com/support>

Account Settings

Support form

Maintenance

Log Out

+ 🌙 🌞

Company name

Prenom / Firstname

Nom / Name

Email*

Sujet du Ticket / Ticket name*

Description du ticket / Ticket description*

Comment reproduire le probleme? / How to reproduce the issue?*

Veuillez copier l'URL correspondant au probleme rencontre et donner les details essentiels afin de le reproduire

Attacher un fichier ou Capture d'ecran / Attached File or Screenshot

Aucun fichier choisi



ATTENTION

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