



EC Crawler Documentation

Single Digital Gateway

Commission européenne, B-1049 Bruxelles / Europese Commissie, B-1049 Brussel - Belgium. Telephone: (32-2) 299 11 11.
Office: 05/45. Telephone: direct line (32-2) 2999659.

Commission européenne, L-2920 Luxembourg. Telephone: (352) 43 01-1.

Table of Contents

- I. Goal..... 2**
 - 1.1. Problem looking to resolve..... 2**
- II. EC crawling process and the metadata 3**
 - 2.1. Notifying a link for Crawling 4**
 - 2.2. Crawling process..... 5**
 - 2.3. Metadata..... 5**
 - 2.3.1. Canonical tag..... 8**
 - 2.4. How can you customize your website crawl process? 10**
- III. Crawler statistics..... 11**
 - 3.1. Problem looking to resolve..... 11**
 - 3.1.1 Usability..... 11**
 - 3.1.2. Administrative..... 12**
 - 3.2. List of Crawled Events..... 13**
 - 3.3. Crawler Event Statistics 15**

I. Goal

The present document has the purpose to explain the **Crawler functionality** currently implemented in the SDG links repository and presents the Crawler's features and the **Crawler Statistics** module that oversees its constant run and would allow to achieve the expected benefits.

1.1. Problem looking to resolve

The first months of experience with the SDG links repository have shown some drawbacks and limitations on the way links can be added into the system.

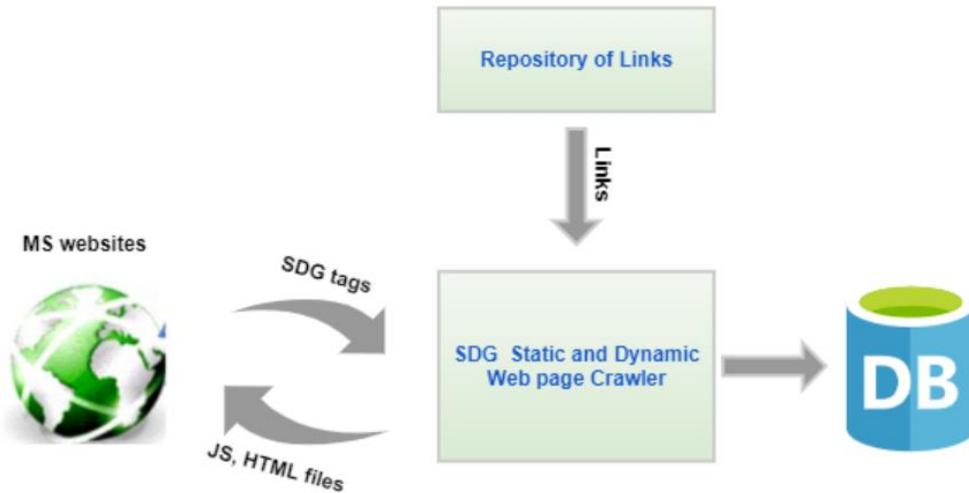
In particular, Member States that had many links to notify needed to make a great effort in adding all the data in the system.

Several proposals have been raised to tackle this, like offering the possibility to import the links into the system:

- By means of file import functionality;
- Having a web service that can ingest links directly from Member States' IT systems.

There was a need for more, and by adding a **Crawler functionality** to the existing links repository we gave Member States the possibility to only notify the web folders containing the links that needed to be added into the system. The Crawler will be in charge of adding/updating/deleting the links from the notified web folders.

Below is a high-level representation of Crawler interaction with Member states websites:



II. EC crawling process and the metadata

In order for the EC crawler to be able to get the required information from the pages that had been notified by the Member States, a number of meta-tags will need to be present in the generated html code of the web pages. We tried to use Dublin Core (<https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>) standard elements whenever suitable.

The crawler is an automated system that takes notified web folders and searches through all the child links' for the ones that contain all the required metadata.

In order to cover all the technologies used to generate websites we have developed two different crawlers:

- the default Crawler which is in charge of all the static web sites
- the JavaScript Crawler for web sites that are generated dynamically using JavaScript.

Note: The **user agent** for the SDG crawler is **EC-SDG-Crawler**.

The crawler **runs continuously** in order to keep the list of metadata links (Web pages) up to date.

* The JavaScript crawler needs manual intervention in order to start. We are trying to run it at regular intervals but if you need updates to your datasets please send us an email and we will run the crawler for your dataset.

2.1. Notifying a link for Crawling

Only Web Folder (parent links) can be notified for crawling. This can be done when a link is added in the repository or when an existing link is edited.

There are two types of Crawlers available, depending on your websites' technical platform. You can choose the appropriate crawler from the Add/Edit link interface:

- ✓ If website contains **static pages**, then the best solution would be to select **“Normal”**
- ✓ If website is built on top of a javascript framework like Angular, React, VueJS, etc., then the best solution would be to select **“JavaScript”**

See image below:

URL : <https://xml2k.co.gov.mt>

Title : **testXML1acAM**

Description : **XML web folder**

Display SDG Dashboard Title/Description in Search Results? :

Uri Type : **Web folder**

National Locations : MALTA (MT) ;

Type : **Information** Procedure

Crawl this URL?: Normal JavaScript

Sitemaps : N/A

Excluded paths : <https://xml2k.co.gov.mt/pe> ; <https://xml2k.co.gov.mt/pe2> ;

Ignore parameters : p1 ; p2 ;

Categories :

Status : **Published**

Owner : **Coordinator National**

Last update : **Mon Dec 18 2023 07:54:38 GMT+0100 (Central European Standard Time)**

Once the websites have been notified, the Crawler functionality will pick them up on the next scheduled Crawl event and details on what was crawled can be seen in Crawler Statistics module.

2.2. Crawling process

- 1) Retrieves the list of all the active countries in the system
- 2) Iterates on the retrieved dataset and start with the first member state (e.g. Austria)
- 3) Retrieves for each member state all the notified links marked for crawling
- 4) Notified links are being crawled **one by one** looking for web pages that contain the following attributes:

- Language (lang / meta.language)
- SDG tag (sdg-tag)
- Country Code (DC.ISO3166)
- Sub-national location (DC..Location)
- Content type (DC.Service)
- Classification information code (policy-code)
- Classification information description (DC.Policy)
- Canonical tag (rel=canonical)

5) After the crawling process has been completed for a notified link, all the valid web pages that have been collected will be stored in the links repository alongside with the collected metadata.

2.3. Metadata

In order for the EC crawler to be able to get the required information from the pages that had been notified by the Member States, a number of meta-tags will need to be present in the generated html code of the web pages.

Below you can find the list of the available **meta-tags**:

Attribute	Attribute name	Mandatory?	Details	Controlled dataset
Language	lang	Yes (either solution is accepted)	It can be found in the following structures: <code><html lang="en"></code>	

	meta.language		<p>Or</p> <p>for the Member States that have don't have properly formatted websites in the metadata section</p> <pre><meta name="DC.language" <u>content="en"</u>></pre>	
sdg tag	sdg-tag	Yes	The sdg-tag will contain the text sdg identifying the page as part of the Single Digital Gateway	{sdg}
Country Code	DC.ISO3166	No	It will contain the two characters ISO 3166-1 representation of names of countries.<meta name="DC.ISO3166" content="BE"/>	The EEA countries
Sub-national location	DC.Location	only if applicable	<p>Describes the NUTS or LAU location id for which the content on the page is valid.</p> <p>https://dublincore.org/specifications/dublin-core/dcmi-terms/#Location</p> <p>It can be found in the following structure:</p> <pre><meta name="DC.Location" content="BE100"/></pre>	<p>The NUTS and LAU location ids</p> <p>The complete NUTS 1-3 and LAU datasets in excel format</p> <p>SDG-nuts-1-3.xlsx</p> <p>EU-28-LAU-2019-NUTS-2016.xlsx</p> <p>The NUTS 1-3 and LAU datasets csv format</p> <p>(to be uploaded soon)</p>
Content type	DC.Service	Yes	It will contain the information about the	{information;

			<p>type of content present on the page</p> <p>https://dublincore.org/specifications/dublin-core/dcmi-terms/#Service</p> <p>It can be found in the following structure:</p> <pre><meta name="DC.Service" content="information"/></pre>	procedure}
Classification information code	policy-code	Yes	<p>It will contain the information about the code of the content area covered by the page according to Annex I and II.</p> <p>Only lowest level categories are accepted (the codes containing the letter of the category and the number of the area)</p> <p>It can be found in the following structure:</p> <pre><meta name="policy-code" content="C2"/></pre> <p>In case the content covers multiple content areas we can put multiple codes in the content attribute of the content-area-code meta tag separated by ","</p>	<p>Annex I and II area codes</p> <p>(only the codes containing the letter of the category and the number of the area)</p> <p>annex I and II.csv</p> <p>annex I and II.xlsx</p>
Classification information description	DC.Policy	No, but recommended	<p>It will contain the information about the name of the content area covered by the page according to Annex I and II</p> <p>https://dublincore.org/specifications/dublin-core/dcmi-terms/#Policy</p> <p>It can be found in the following structure:</p> <pre><meta name="DC.Policy" content="acquiring and renewing a driving licence"/></pre>	<p>Annex I and II area names</p> <p>annex I and II.csv</p> <p>annex I and II.xlsx</p>
Canonical tag	rel=canonical	only if applicable	The canonical tag is recommended to be used in the web pages that will need to be	

			<p>crawled to make sure we are not registering duplicated pages.</p> <p>It can be found in the following structure:</p> <pre><link rel="canonical" href= "https://verwaltung.bund.de/leistungsverzeichnis/EN/leistung/BB/102088827" /></pre>	
--	--	--	---	--

*if one of the tags needs to contain more than one value than these values need to be separated by a semicolon sign (;).

For example:

```
<meta name="DC.Service" content="information;procedure"/>
```

In order for the crawler to be able to find the right meta tags and extract the information from the web pages the DC prefix needs to be used whenever the meta tag name is part of the Dublin Core elements.

The servers that host the websites notified in the Links repository for crawling will need to allow access to the crawler.

2.3.1. Canonical tag

The canonical tag is recommended to be used on the web pages that will need to be crawled to make sure we are not registering duplicated pages. For example, on Your Europe we are using filters to display the relevant contact points depending from which section the users arrive to the contact points pages. https://europa.eu/youreurope/citizens/national-contact-points/ireland/index_en.htm?contacts=65891 and https://europa.eu/youreurope/citizens/national-contact-points/ireland/index_en.htm?contacts=65641 are both instances of the original page https://europa.eu/youreurope/citizens/national-contact-points/ireland/index_en.htm but

display filtered content depending on the users' journey towards the contact points page. The html content of all three pages is identical although we have here three unique URLs.

In order for the commercial search engine to understand which page needs to be indexed, we are using the canonical tag with the value *https://europa.eu/youreurope/citizens/national-contact-points/ireland/index_en.htm*. The other two pages are just different presentations of the same content.

The EC crawler will use the canonical URL in a similar fashion to exclude saving in the repository duplicated pages.

- **Tagging Example 1**

If we consider the Belgian portal <https://www.belgium.be> and we want to tag the information about acquiring and renewing your driving licence (C2 in Annex I) the tagging will look like the example below. For this theoretical exercise, we assumed that the information on the page only applies for the Brussels region (BE1).

Web Page: https://www.belgium.be/fr/mobilite/permis_de_conduire/obtenir_un_nouveau_permis

```
<html lang="fr"/>
<meta name="sdg-tag" content="sdg"/>
<meta name="DC.ISO3166" content="BE"/>
<meta name="DC.Location" content="BE1"/>
<meta name="DC.Service" content="information"/>
<meta name="policy-code" content="C2"/>
<meta name="DC.Policy" content="acquiring and renewing a driving licence"/>
```

- **Tagging Example 2**

If we consider the Belgian portal <https://www.belgium.be> and we want to tag the information and the procedure about opening a business (J1 in Annex I and W1 in Annex II) the tagging will look like the example below. For this theoretical exercise, we assumed that the information on

the page only applies for the Brussels region (BE1) and on the page, we have information and the procedure.

Web Page : https://www.belgium.be/fr/economie/entreprise/creation/etapes_principales

```
<html lang="fr"/>
<meta name="sdg-tag" content="sdg"/>
<meta name="DC.ISO3166" content="BE"/>
<meta name="DC.Location" content="BE1"/>
<meta name="DC.Service" content="information;procedure"/>
<meta name="policy-code" content="J1;W1"/>
<meta name="DC.Policy" content=" registering, changing the legal form of or closing a
business (registration procedures and legal forms for carrying out business);General registration
of business activity, excluding procedures concerning the constitution of companies or firms
within the meaning of the second paragraph of Article 54 TFEU"/>
```

*please note that there may not be an exact match between the example above and the specific situations in your country. Therefore, the instance illustrated above is only for exemplification purposes.

2.4. How can you customize your website crawl process?

You can tailor the metadata extraction by completing various optional fields, including:

Excluded Paths: These are specific paths within a website that you choose to exclude from the crawling process. By specifying excluded paths, you instruct the crawler to avoid certain sections of the website, which can be useful for focusing the crawl on relevant content and avoiding unnecessary or sensitive areas.

Ignored Parameters: Ignored parameters are parameters in a URL that you choose to disregard during the crawling process. Crawlers often encounter URLs with dynamic parameters that don't

affect the content but can lead to unnecessary duplication in the crawl results. By specifying ignored parameters, you optimize the crawling process by excluding irrelevant variations of URLs.

Sitemaps: The crawling process will be more optimal by incorporating them, offering the crawler the exact pages it needs to visit without having to read the whole dataset of URLs present in a website, some of which might be found multiple times over the numerous pages of a website. This information aids crawlers in efficiently reading all the individual pages, ultimately leading to a more focused and effective crawl.

In summary, excluded paths and ignored parameters allow you to fine-tune the crawling process by excluding specific content or URL variations, while sitemaps, with their optional fields, provide additional information to guide the crawler in understanding the structure and content of your website.

III. Crawler statistics

3.1. Problem looking to resolve

The Crawler functionality is a service that runs alongside the Repository of Links and constantly ingest links, without having a tool to have an overall picture of the crawling process.

Several problems have been identified:

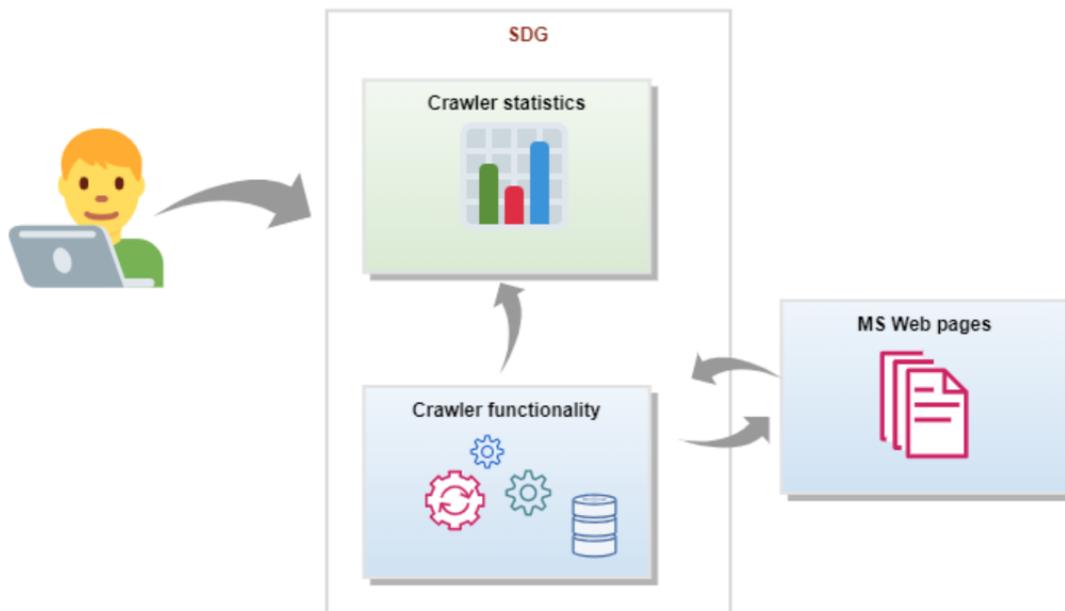
3.1.1 Usability

- Statistics about the crawler activity could be gathered only from logs
- Issues and problems were again collected from logs
- Member states had no access to crawler stats or information

3.1.2. Administrative

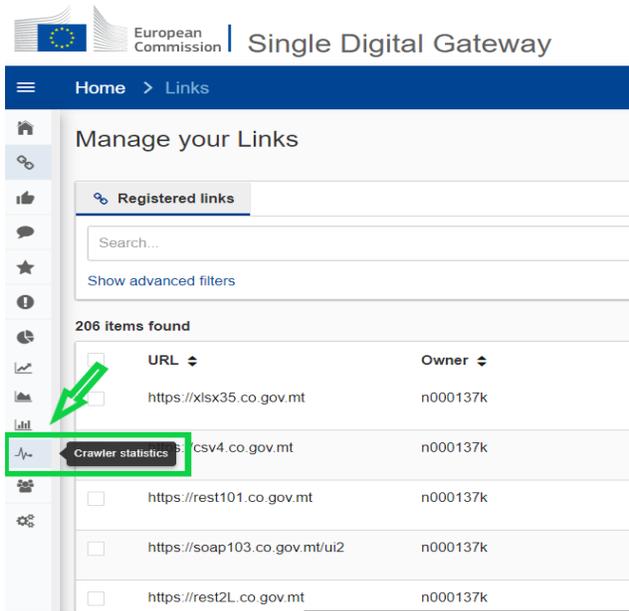
- Not knowing when crawler was running or blocked
- No control over the crawler lifecycle

The solution was to implement an interface that would tackle all the issues and concerns raised and would give the application user more control over this important functionality.



Crawler Statistics module is located on the left navigation bar by expanding **Statistics** module and after clicking on '**Crawler Statistics**' icon.

See image below:



This module is available for all national users as follows:

- **National Coordinators / National Observers** - have access to all the information specific for their Member state.
- **National Service Providers** - have access to information on the links for which they are responsible.

3.2. List of Crawled Events

After clicking on 'Crawler statistics' icon, the user will see the 'List of Crawled Events' page, where all present and past crawl events are listed, for each of them some information being available in a table format:

- ✚ **Crawler ID** - which is the unique ID assigned to each event.
- ✚ **Status** – represents the status of the crawler event, which can be:
 - **'In Progress'** -> crawler is still processing other countries; the overall crawler has not yet finished, there are countries that are pending to be crawled.
 - **'Completed'** -> crawling process has been finished.
 - **'Interrupted'** -> occurs very rarely, but what it means is that for some reasons the Crawler event has been interrupted for a short amount of time.

Note: If value 'Pending' is present per event then the data is currently being processed and there is no available info yet.

 **Time duration** - which tells us how much time this event took.

 **Last update** - the time when the information was last updated for the crawl event.

See image below:

List of Crawled Events

Please select a range of dates

07/09/2022 - 07/10/2022

22 items found

Crawler ID	Status	Time duration	Last update	ACTIONS
20221007	In progress	00:52:58	2022-10-07 09:41:12	
20221006	Completed	00:52:32	2022-10-06 09:44:45	
20221005	Completed	00:52:36	2022-10-05 09:40:40	
20221004	Completed	00:52:50	2022-10-04 09:44:36	
20221001	Completed	00:35:48	2022-10-03 09:01:31	

1 2 3 5

Note: JavaScript events are highlighted in blue while the crawler ID has a dedicated pattern containing JS label at the end.

Example below:

20220905 (JS)	1	Completed	00:54:10	2022-09-06 20:59:33	
---------------	---	-----------	----------	---------------------	---

In the left corner of the list of items page, the user has also the possibility to filter crawl events by a specific interval of time:

See image below:

List of Crawled Events

Please select a range of dates

08/09/2022 - 08/10/2022

3.3. Crawler Event Statistics

If the user selects a crawl event by clicking the View icon , it will then be redirected to the 'Crawler Event Statistics' page, which is split into three tabs:

- General
- Critical issues
- Medium issues

Note: In order to have the latest information about the Crawler it is recommended to select one of the recent ones.

➤ General tab

On the 'General tab' all notified links that have been marked for Crawl will be listed, each of them coming with some extra information attached such as:

- ✚ *Pages Crawled* - which shows the total amount of metadata links visited by the Crawler for that notified link.
- ✚ *Pages with metadata* - where two information are displayed, the first being the 'Total amount of metadata links found' while the other is the 'Percentage of metadata links found' out of the total visited links.
- ✚ *Time duration* - which represents the total time it took for the notified link to be crawled.
- ✚ *Last update* - the time when the information was last updated for this notified link
- Actions* - there is a log that can be downloaded where the user can see the order and path the Crawler took on the website in order to find metadata links.

See image below:

Crawler Event Statistics

General Critical issues **76** Medium issues **9**

78 items found

URL ↕	Pages crawled ↕	Pages with metadata ↕	Time duration ↕	Last update ↕	ACTIONS
https://ncpe.gov.mt/en/Pages/Comments_Form.aspx	0	0 (0 %)	00:00:40	2022-10-06 08:52:53	↓
https://pulizija.gov.mt/mt/services/Pages/Emergency-Services.aspx	1	1 (100 %)	00:00:40	2022-10-06 08:53:33	↓
https://businessenhance.gov.mt/SubMenus.aspx?EEOCdQZCR0JJOK8qqHch9qO8 PRFL9k	1	1 (100 %)	00:00:40	2022-10-06 08:54:13	↓
https://customs.gov.mt/bus/introduction-to-taric	1	1 (100 %)	00:00:40	2022-10-06 08:54:54	↓
https://ncpe.gov.mt/en/Pages/Rights_and_Obligations/Equality-in-Goods-and-Services.aspx	1	1 (100 %)	00:00:40	2022-10-06 08:55:34	↓
	80	65 (5650 %)	00:52:25		

1 2 3 5

At the bottom of the table, highlighted in blue, the total values for each column are displayed.

➤ Critical issues tab

This tab contains a list with all the notified websites that are considered as having critical issues, highlighted with **red**.

The criteria for a link to be listed in this tab are as follows:

- ⇒ If a Website has less than **two pages** visited;
- ⇒ If a website takes more than **3h 30 min** to crawl;
- ⇒ If the number of saved pages with metadata is less than **10%** from the total crawled pages.

Critical information table contains:

- ✚ *URL* – notified link marked for JS or regular crawling;
- ✚ *Crawled?* – evaluating if the link has been crawled or not.

Possible options:

- Yes (if we have more than 2 pages visited) or,
- No (having less than 2 pages visited);
- ✚ *Pages with Metadata* – the total number of metadata links found for the notified parent link;
- ✚ *Time Duration* – the total time it took to crawl the notified parent website;

- ✚ **Actions** – the user can download the log file used to debug the crawling process for a parent website.

See image below:

Crawler Event Statistics

General **Critical issues 76** Medium issues 0

URL ↕	Crawled ? ↕	Pages with Metadata ↕	Time duration ↕	ACTIONS
https://ncpe.gov.mt/en/Pages/Comments_Form.aspx	No	0 (0 %)	00:00:40	↓
https://pulizija.gov.mt/services/Pages/Emergency-Services.aspx	No	1 (100 %)	00:00:40	↓
https://businessenhance.gov.mt/SubMenus.aspx?E0CdQZCR0JJOK8qqHch9qQ8PRFL9k	No	1 (100 %)	00:00:40	↓
https://customs.gov.mt/bus/an-introduction-to-tanic	No	1 (100 %)	00:00:40	↓
https://ncpe.gov.mt/en/Pages/Rights_and_Obligations/Equality-In-Goods-and-Services.aspx	No	1 (100 %)	00:00:40	↓

1 2 3 5

➤ Medium issues tab

'Medium issues tab' - contains a list with all notified websites that considered as having minor issues highlighted with **orange**.

The criteria for a link to be listed here are as follows:

- ⇒ If a Website takes more than **2h** to be crawled.
- ⇒ If the number of saved pages with metadata is less than **30%** from the total crawled pages

The information presented in the Medium issues tab is:

- ✚ **URL** – notified link marked for JS or regular crawling
- ✚ **Pages with Metadata** – the total number of metadata links found for the notified parent link
- ✚ **Time Duration** – the total time it took to crawl the notified parent website

Actions – the user can download the log file used to debug the crawling process for a parent website.

See image below:

The screenshot shows a web interface titled "Crawler Event Statistics". At the top, there are three tabs: "General", "Critical issues 152", and "Medium issues 4". The "Medium issues 4" tab is selected and highlighted with a yellow border. Below the tabs is a table with four columns: "URL", "Pages with Metadata", "Time duration", and "ACTIONS". The table contains four rows of data. At the bottom right of the table, there is a pagination control showing "1" of "5" items.

URL	Pages with Metadata	Time duration	ACTIONS
https://www.prv.se/en/	56 (10.67 %)	00:02:20	
https://www.elsakerhetsverket.se/	55 (13.72 %)	00:03:50	
https://pts.se/en/english-b/	6 (10.53 %)	00:00:50	
https://www.av.se/en	62 (24.41 %)	00:01:30	