To paraphrase Donella Meadows, we can’t impose our will on Internet.
We can listen to what Internet tells us, and discover how its properties and our values
can work together to bring forth something much better than could ever be produced by our will alone.

The Observatory Tool

Getting Started Guide

(Vers. 05 - December 2016)
Index

The Observatory Tool

Introduction

Description

Access

Content harvesting

Web content syndication

Web harvesting

Social media harvesting

Content filtering

Content analysis

Content evaluation

User dashboard

Collaborative features

Exportation of content (Open Data)

Exportation of items

Exportation of sources

API access
1. Introduction

GIPO Observatory is mainly intended to help the global community, including the GIPO team members and its advisory board, with the communication, discovery and analysis of valuable information (i.e. news, events, policy reports, etc.) related with Internet policy.

1.1. Description

The heart of the Observatory is a web application (The Observatory Tool, or GIPO Tool). The philosophy behind it is to scan online quality content sources (compiled and managed by GIPO-Authorized users), and import only those content items that comply with a set of rules (filters) associated with each source.

All users can review each item and share it. The GIPO Tool assigns the metadata suitable for every content item. This process is aided by machine classification and relevance evaluation, and complemented with commenting functionalities to foster collaboration.

1.2. Access

The Observatory Tool is accessible from Internet\(^1\) using any web browser. You can visit the [giponet.org](http://giponet.org) portal where a specific button and/or banner will direct users to the tool.

---

\(^1\) Access the Observatory Tool from here: [http://observatory.giponet.org/](http://observatory.giponet.org/)
2. Content harvesting

The Observatory Tool can manage three primary types of content sources for their automatic processing: web content syndication, web harvesting, and social media harvesting.

For web content syndication the tool relies on RSS and Atom formats. For web harvesting GIPO Tool makes use of Scrapy\(^2\) web crawler. For social media harvesting the tool makes use of APIs offered by the corresponding platform (whenever available and in an open\(^3\) manner).

2.1. Web content syndication

Much Internet content such as news headlines, search results, etc. are shown as lists/tables, or very much resemble them.

---

\(^2\) Scrapy, a fast and powerful scraping and web crawling framework. [https://scrapy.org/](https://scrapy.org/)

\(^3\) Meaning by ‘open’ a free and public access to the API. This may imply some usage restrictions usually based in limiting the number of calls during a period of time, etc.
RSS⁴ is an XML-based format intended for use by computers that allows the syndication of content.

Currently many Web sites make a feed, or channel, available, just like any other file or resource on the server.

A feed contains a list of items or entries, each of which is identified by a link. Each item can have any amount of other metadata associated with it as well. Computers can regularly fetch the feed to get the most recent items on the list.

The most basic metadata for an entry includes a title for the link and a description of it, and additionally metadata associated with it like a category or tags, and other useful information.

Using RSS/Atom feeds is a straightforward and easy option for Web site tracking, because the feed is machine-readable so The Observatory Tool doesn’t have to figure out which parts of the site are important and which parts are just the navigation and presentation.

We want to emphasize that this tool does not act as a mere RSS reader, because data analysis’ procedures are applied to the content retrieved from the feeds.

Note also that you may encounter some RSS feeds that does not provide the referral link to the information in the original source, but a link to the container element in the web page that displays the information. This usually happens with RSS feeds from news aggregators or content discovery tools such as Scoop.it⁵ and the like.

---

⁴ RSS 2.0 specification: [http://validator.w3.org/feed/docs/rss2.html](http://validator.w3.org/feed/docs/rss2.html)

⁵ Scoop.it!: [http://www.scoop.it/](http://www.scoop.it/)
2.2. Web harvesting

Also known as “web scraping” this is a technique of extracting information from websites, especially from those that do not offer any machine-friendly way (such as RSS feeds or APIs) of extracting its content.

For this task GIPO uses a web crawler, that is a software application that scans every page of the website associated with the URL domain of the source, extracting page titles, descriptions, keywords, and other valuable metadata.

2.3. Social media harvesting

The Observatory Tool gathers content from APIs provided by social media platforms such as Twitter\(^6\) and Google Plus\(^7\).

The GIPO Tool accesses to the global stream of public tweet data flowing through Twitter and retrieves only those Tweets that comply with these rules:

- contains any of the terms included in a list of phrases (keywords or hashtags);
- contains at least one web address (URL) not pointing to an image.

Additionally the GIPO Tool features antispam control and detection of duplicate items.

3. Content filtering

Not any content items retrieved from a given source may be suitable for displaying in The Observatory Tool. This usually happens when the content is not closely related with any of the multiple Internet governance topics, or when the most part of the content is marketing information from consulting firms or infrastructure vendors, etc.

One of the biggest advantages of this tool is its ability to automatically discard the non-relevant content items and collect only the interesting ones. What’s more, the tool estimates the relevance of the items and displays them ordered (see ‘Content analysis’ further below). The Observatory Tool also enriches the contents’ metadata by automatically assigning some appropriate tags to every item.

The automatic classification of desired items, for instance under a specific Internet governance issue, is done by means of filters. Unless a GIPO-Authorized user creates at least one filter, no content item will be classified and thus assigned the appropriate tags.

---

\(^6\) Twitter’s Streaming APIs: [https://dev.twitter.com/streaming/overview](https://dev.twitter.com/streaming/overview)

\(^7\) Google+ API: [https://developers.google.com/+/api/](https://developers.google.com/+/api/)
When a GIPO-Authorized user defines a filter for use with any of the sources, it is necessary to set up various parameters so the filter can act as expected. GIPO-Authorized users can add as many filters (of any type and language) as they wish. Filters, as well as sources, can be enabled or disabled individually.

4. Content analysis

Sources are regularly checked and content items are automatically detected according to the filters’ configuration. The data store and any content page within GIPO tool is updated accordingly with all the new items harvested.

In GIPO Tool a first initial set of tags are automatically annotated to items by using semantic analysis. As this is difficult to perform on one’s own, this is currently achieved by using Thomson-Reuters Open Calais® Web Service API. This Web Service allows to automatically annotate content with rich semantic metadata, including entities such as people and companies and events and facts. OpenCalais automatically identifies the language of submitted text and applies a Language

8 Open Calais: http://new.opencalais.com/about/
Identification module before processing the text for entities, events and facts. Today it supports content in English, French and Spanish.

Currently the GIPO Tool can classify an item under one or more codes within these groups:

- **Issues**: the main information fields covered by the content of the item.

  Note that ‘Issues’ is a key taxonomy in GIPO Tool so any selection should be carefully made, as it will have an impact in future end-user searches’ effectiveness.

- **Tags**: other useful keywords.

- **World regions**: countries and regions mentioned most often in the item.

- **Type**: the nature or information structure of the item (blog, event, case, social media, etc.).

The review of items is made from the search results page:

![Search results page](https://observatory.giponet.org)

This view shows a paginated list of items ordered by the relevance of each item. To calculate the relevance of each item the tool uses a scoring formula with these main factors:

- the term frequency - the more times a search term appears in an item, the higher the score
- inverse item frequency - matches on rarer terms count more than matches on common terms
GIPO Tool - Getting Started Guide

- multiple terms match - if there are multiple terms in a query, the more terms that match, the higher the score
- length - matches on a smaller field score higher than matches on a larger field

The right section of this page allows users to filter the list of items by applying multiple filters (facets) to more quickly find desired results.

The facets currently available are: **issues, tags, world regions, and type.**

These are the operations that **all users can perform with the items** collected:

- **Search** for a specific desired item or a group of items that meet users’ criteria.

The tool helps users by offering search history results that are based on user's own previous searches.

As you type a list of previous search terms, sharing the same first characters, will automatically appear in the dropdown menu at the bottom. To select any of them just click on it, or use the down key to navigate to it, and then hit Enter.

You don't have to know advanced searching techniques to find your items in GIPO Tool, because searching can be as simple as typing a few terms in the search box. Additionally, you can refine search results by using the faceted navigation. But if you want to get more specific search results, you can use these tips that can be quite helpful depending on what you’re searching for:

**Boolean operators**

<table>
<thead>
<tr>
<th>Operator</th>
<th>How to use it</th>
</tr>
</thead>
</table>
| “        | Requires words in quotes to be searched as a phrase, in the same order.  
*Example: “net neutrality”* |
| - (minus)| Excludes single words or phrases (words between quotes) immediately following it. |
**Example: “net neutrality” - “free basics”**

**OR**

Searches for either of the words or phrases (words between quotes). If there is no operator between two search terms, the OR operator is used automatically.

*Example: “net neutrality” OR cybersecurity*

**AND**

Searches for items where both words or phases exist (anywhere in the item).

*Example: cyberbullying AND “cell phones”*

**NOT**

Items that contain the search term after the NOT are not included in the search.

*Example: “human rights” NOT “financial crisis”*

---

**Fuzzy / Wildcard searching**

<table>
<thead>
<tr>
<th>Character</th>
<th>How to use it</th>
</tr>
</thead>
</table>
| ? (question mark) | Items that contain search terms variations that match a single character.  
*Example: cent?? cent centre OR center OR centro OR centri OR...* |
| * (asterisk) | Items that contain search terms variations that match zero or more sequential characters.  
*Example: cent* cent OR cents OR center OR central OR century OR...* |

---

**Fuzzy / edit-distance searching**

You can use the tilted symbol ( ~ ) to find terms that are close to the original, or misspellings.

*Example: Europa~ Europol OR Euroopa OR Europe OR European OR euro OR...*  

**Proximity searching**

You can use the tilted symbol followed by a number (~N) to find first term within a number N of positions of second term.

*Example: “open data” ~2 “open data” OR “open government data” OR “open sharing of data”*

---

**Multi-language searching**

You can use lang:LANG-ISO-CODE to search for items written in a given language represented by its ISO 639-1 code⁹, e.g. en, de, es

---

Multi-tag searching
You can use tags:TERM to find all items tagged with some specific terms.
Example: tags:(igf OR "internet governance forum")

Date-based searching
You can use tstamp:DATE-BASED-EXPRESSION to find all items within a period relative to fixed moments in time. The "NOW" parameter is used to set a fixed time for evaluating date based expressions. All date based expressions are evaluated relative to the UTC Time Zone.
Example: tstamp:[NOW-5DAY TO NOW] -tweeted

- **Browse** the list of Internet governance issues to narrow the set of items listed.

- **View** the information of any item from any source.

- **Comment** and **share** any item (see next section for more information about this).

5. **Content evaluation**

The ability of all users to provide feedback, and share items is an important feature of the GIPO Tool, by including appropriate modules that will help users evaluate the quality of the content within the tool.
Moreover these feedback enable the discovery of high-quality and trending content, as well as improves content filtering, and recommendation that may also become a mean of interaction between users in GIPO Tool.

5.1. User dashboard

The home page of the GIPO Tool displays a compact view of some selected content that may be interesting for the user. Among this content there are lists of recently harvested items, user-defined searches, graphs representing trends in topics or geographic areas, etc.

The “Latest items” list shows the last ten items collected by the platform:

A fully customizable graphical dashboard is also available. More information about it is available in a different document (the GIPO Tool Dashboard manual).
5.2. Collaborative features

All users have the option to include their own commentaries in sources and filters, plus share any item by email or on whatever social sharing service in which the user is registered (among Facebook, Twitter, Google+, or LinkedIn).

6. Exportation of content (Open Data)

Open data is data that anyone can access, use or share. The GIPO Tool offers by default output feeds in open, machine-readable and standard format.
6.1. Exportation of items

Whenever a GIPO collects an item, its metadata is routed to the appropriate output feed (in XML-RSS\(^{10}\) open format). This way items exported from the tool can be automatically consumed by other systems, such as web content managers, email marketing platforms, or even desktop and online applications such as MS Excel, Google Sheets, mobile apps, etc.

Every output feed in GIPO contains a list of the items or entries gathered about a specific Internet governance issue covered in the tool. Each entry in the feed is identified by a link and has some additional metadata, like a title for the link, a description, the category or tags, and other useful information.

Computers can regularly fetch the feed to get the most recent items. A feed element consists of a start tag and end tag, with the content inserted in between:

\[
\text{<tagname>Content goes here...</tagname>}
\]

\[
\text{<item>}
\]

An item element represents distinct content published in the feed such as a news article, weblog entry, etc.

An item contains the following child elements: title, link, description, pubDate, dc:rights, author, source, category. These elements are not present more than once in an item, with the exception of category.

\[
\text{<title>}
\]

An item's title element holds character data that provides the item's headline.

\[
\text{Example: <title>Community, Identity, Speech, and Power</title>}
\]

\(^{10}\) [http://validator.w3.org/feed/docs/rss2.html](http://validator.w3.org/feed/docs/rss2.html)
<link>
An item's link element identifies the URL of a web page associated with the item.

Example: <link>https://cdt.org/blog/community-identity-speech-and-power/</link>

<description>
An item's description element holds character data that contains a summary of its contents.

Example: <description>A lot has changed since this introduction to CDT on our first website, circa 1994. The web has grown to contain more than 4 billion pages on an underlying infrastructure that connects nearly half...</description>

<pubDate>
An item's pubDate element indicates the publication date and time of the item.

Example: <pubDate>Wed, 27 Jun 2016 15:22:33 +0000</pubDate>

<dc:rights>
An item’s Dublin Core rights element identifies the URL of a web page associated with the license under which the item is made available.


<author>
An item's author element provides the name of the person who wrote the item.

Example: <author>Nuala O’Connor</author>

<source>
An item's source element indicates the fact that the item has been republished from another source. The value of the source is the title of the source as listed in GIPO (http://observatory.giponet.org/sources).

Example: <source url='https://cdt.org/blog/feed/'>Center for Democracy & Technology - Blog Posts</source>

<category>
An item’s category element identifies a category to which the item belongs. This element includes a domain attribute that identifies the category’s taxonomy.

Example: <category domain='http://observatory.giponet.org/aspects'>Human Rights</category>

---

11 The Dublin Core Metadata Initiative http://dublincore.org/
12 GIPO Tool classifies internet governance issues on the basis of the taxonomy developed by Diplo Foundation.
Tip: if you are a regular user of Google Sheets you can try the IMPORTFEED\textsuperscript{13} or IMPORTXML\textsuperscript{14} functions to import GIPO’s structured data.

6.2. Exportation of sources

All sources in GIPO tool can be downloaded as an Outline Processor Markup Language (OPML) file. The OPML\textsuperscript{15} is an XML-based open format that has become popular as a method for exchanging subscription lists between feed readers and aggregators.

Tip: you can use the ‘import’ feature included in many desktop or online feed readers to add all the sources covered by GIPO tool.

6.3. API access

GIPO tool offers a REST-like API to programatically access the search server. Users can query it via HTTP GET and receive results in JavaScript Object Notation (JSON), XML, or CSV format. Queries are executed by creating a URL that contains all the appropriate query parameters (although some client applications may hide this detail to users).

The fully-qualified URL must have the following format:

\begin{verbatim}
http://observatory.giponet.org/api/solr-proxy/rest/proxy/select?parameterkeyvalue
\end{verbatim}

Where parameterkeyvalue can be one or more parameter key and value pairs indicating the details of your query. To construct a parameter key and value pair, you must concatenate each parameter key with an equal sign (=) and a value string. For example: q=broadband. To create a string of parameter key and value pairs, you must concatenate each pair using an ampersand (&). For example: \texttt{q=broadband&rows=100}

\textsuperscript{13} https://support.google.com/docs/answer/3093337
\textsuperscript{14} https://support.google.com/docs/answer/3093342
\textsuperscript{15} http://dev.opml.org/spec2.html