

WF DEVELOPMENT

Introduction and guidelines

February 2018

Unauthorized use, duplication, or modification of this document in whole or in part without the written consent of Verint Systems Inc. is strictly prohibited. By providing this document, Verint Systems Inc. is not making any representations regarding the correctness or completeness of its contents and reserves the right to alter this document at any time without notice. Features listed in this document are subject to change. Contact your Verint representative for current product features and specifications. All marks referenced herein with the ® or TM symbol are registered trademarks or trademarks of Verint Systems Inc. or its subsidiaries. All rights reserved.   
All other marks are trademarks of their respective owners.

© 2018 Verint Systems Inc. All rights reserved worldwide.

Table of Contents

[1.1 Acronyms 1](#_Toc506821104)

[1.2 Webint Collect High-level Architecture 2](#_Toc506821105)

[1.3 WebFlows 3](#_Toc506821106)

[Definition 3](#_Toc506821107)

[Prerequisites 3](#_Toc506821108)

[1.4 WebFlow Entities 4](#_Toc506821109)

[1.5 The WebFlow Builder 8](#_Toc506821110)

[The Birdeye Panel Menu 9](#_Toc506821111)

[The Browser Panel Menu 9](#_Toc506821112)

[Creating New Steps 11](#_Toc506821113)

[1.6 Executing WebFlows 13](#_Toc506821114)

[Create Web Platforms 13](#_Toc506821115)

[Publish WebFlow 15](#_Toc506821116)

[Attach WebFlow to Web Platform 16](#_Toc506821117)

[Create a Virtual Agent 16](#_Toc506821118)

[Create a Collection Request 19](#_Toc506821119)

[1.7 WebFlows’ Execution Info 22](#_Toc506821120)

* 1. Acronyms

|  |  |
| --- | --- |
| **Term** | **Description** |
| WebFlow (WF) | A series of web actions that are being executed according to defined sequence |
| Super WebFlow (SWF) | A combination that defines interactions between more than one WF |
| Entity | Tabular object that is being utilize to store information within WFs |
| Crawler Unit (CU) | A component that serves as WebFlows executer. |
| Custom Action (CA) | Allows the user to create JS functionality that can be applied within WebFlows. |
| Virtual Agent (VA) | Web site login credentials of an existing user which used to access the relevant content |
|  |  |

* 1. Webint Collect High-level Architecture

The Webint Collect is structured from five main components:

* **Webint Collect Admin** – Provides a UI tool to build, execute and monitor WebFlows.   
  In addition it manages the Focal Collect configurations.
* **Webint Collect Site Manager** – Responsible for distributing tasks between the
* **Webint Collect Crawler Unit** – Provides a platform for executing the WebFlows.
* **Webint Collect Distributer** – Responsible for exporting collected data into files.

*Webint Collect components diagram*



Admin



Site Manager



Distributer

Activate SM with the CR

Notify for   
availability

Assign tasks



Crawler Units



Notify upon start/end of Job



SQL Server

Export rows from DB

Write collected rows to DB

* 1. WebFlows

###### Definition

A WebFlow is a set of specific actions that can be executed over various web interfaces in a predefined sequence.

###### Prerequisites

In order to create new WebFlow the following prerequisites should be taken into consideration:

* **Entities**  
  Relevant input, output and runtime entities should be defined.  
  This topic will be explained in more details in later part of this document.
* **Virtual Agents**If the specific site requires login, it is recommended to create the Virtual Agent in the advance.
* **Collection Plan**Before starting to build the WF, one should come up with a collection plan that illustrates the planned flow of collection. This step of design will assist to build a consistent WF that match the various challenges of the specific web site and will create the development process much more efficient.

* 1. WebFlow Entities

WebFlow entities are used to store data records for various purpose.  
The entities are divided into three categories:

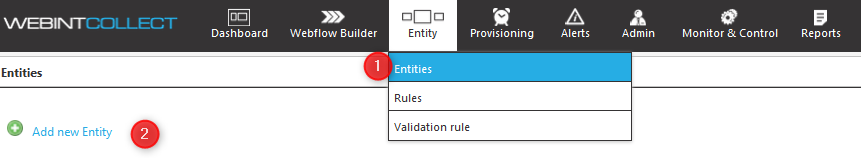
* **Input Entity**  
  Used for passing parameters to the WebFlow (such as: target URL, keyword and so on).
* **Runtime Entity**   
  Used for manipulation of parameters on the WF runtime.
* **Output entity** (Data entity)  
  Used to store the collected data.

***Creating new entity***

In order to create new Entity, follow the steps below:

**Step 1**

At the upper menu click **Entities** under the **Entity** **Menu** and then **Add new Entity**.

****

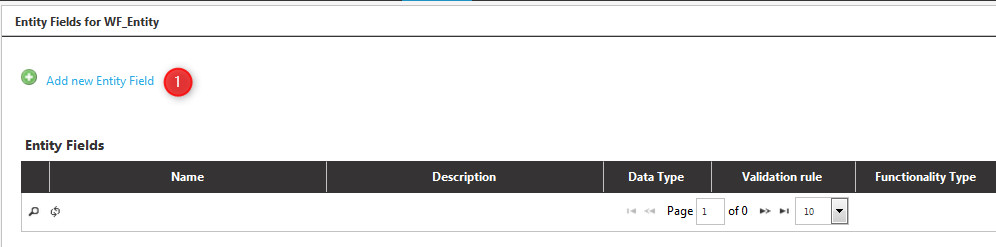
**Step 2**

Type the Entity’s **Name** and **Description**. Then click **Save**.

****

**Step 3**

Click **Add new Entity Field**.



**Step 4**

In the **New Entity Field Editor** enter the following:

* **Name –** The field name.
* **Data Type** – Select which data type should be used. For input entities’ main input (keyword/URL), string array should be selected to allow to send a list of values.   
  For output entities it is recommended to use string (50/250/4000/long) data types according to the relevant expected string length.
* **Label** – used only for input entities. Will determine the caption of the field in Webint Analytics UI.
* **Functionality Type** – used only for input entities. The type will determine what type of collection request will be used. The common used functionalities are:
  + **Username –** store the Virtual Agent’s username.
  + **Password –** store the Virtual Agent’s password.
  + **Search\_Keyword\_Profile –** used to collect searched web profiles by keyword.
  + **Search\_Keyword\_Activities –** used to collect searched activities (posts, images, etc.) by keyword.
  + **Collect\_URI\_Activity –** used to collect specific activity (post, image, etc.) by keyword.
  + **Collect\_URI\_Profile –** used to collect specific web profiles by URL.
* **Description** – used only for input entities. Will determine the tooltip’s caption of the field in Webint Analytics UI.
* **Order** – used only for input entities. Will determine the relative position of the field compared with the rest of the entity’s fields in Webint Analytics UI.
* **Validation Rule** – used only for input entities. Will determine the regular expression (regex) validation rule that will be used to validate the entered value.
* **Renderer** - used only for input entities. Will determine the input field format that will be presented in in Webint Analytics UI, text/check/combo box, date picker, etc.
* **Visible** - used only for input entities. Will determine if the field will be visible in Webint Analytics UI.
* **Global** – when checked, the filed can be shared between different WebFlows inside a SuperWebFlow.
* **Use with WCS –** used to determine which fields should work for enumeration of records.
* Then click **Save**.



**Step 5**

Repeat step 3 and 4 for each field of the entity.

* 1. The WebFlow Builder

This is the WebFlow developer interface. Within this page, the developer utilizes the different tools in order to create a custom WebFlow.

The UI is divided into 4 main parts:

* **Birdeye panel** – Contains all the WebFlow’s steps by the given order and provides a set of management tools.
* **Action panel** – Displays the properties for each selected step at the **Birdeye panel**.
* **Browser Panel** - Displays the current loaded page, enable to create different steps based on the page elements selection.
* **Selection Panel** – Displays the properties of the current/selected HTML tag and allows to insert manually specific X-path commands.



###### The Birdeye Panel Menu

Enables the user to perform actions within the birdeye panel.

* **Previous Step**  - Navigate to previous step.
* **Next Step**  – Navigate to next step.
* **Delete**  - Delete selected step.
* **Apply Changes**  – Save WebFlow.
* **Save and Close**  – Save WebFlow and escape the WebFlow builder UI.
* **Recache Flow**  – Loads the WebFlow pages and re-cache it.
* **Run Flow Now**  – Execute the WebFlow instantly.

###### The Browser Panel Menu

Enables the user to perform actions on the current loaded page.

**Navigate Menu:**

* **Click** – Perform click action on selected clickable web objects.
* **For Each** – Allows iteration over selected list of HTML tags.
* **End.. For Each** – Virtual step to define the end of a **For Each**. After clicking this option, next steps will be created outside the **For Each**.
* **GoTo next page** – When looping on a list of results, sometimes there are more than one page to show the results. The step **GoTo Next Page** must be mapped to the link of the next page. After you select it, the report builder will request the step where you want to return after clicking the **Next** link (generally will be the **For Each**).
* **Condition** – Split the execution of the WebFlow according to the existence or not of a specific element.
* **Condition Parameter** – Allows navigation according to parameter comparison.
* **Load URL** – Load given URL.
* **Load List URLs** – Load given list of URLs.
* **Press Enter** – Mimic pressing the Enter key.
* **Repeated click** – Used for elements executing AJAX requests, it is recommended to use this step and not the regular **Click**.
* **Scroll to bottom** – Mimic mouse scroll down within the loaded page.
* **Decaptcha** – During **Logins** may appear **Captcha** images. Using this step, the user will select the corresponding image and input box where to store the resultant value, and the system will fill this element during the execution.
* **While –** Allow to perform “while” loop over a course of steps. The “while” condition is based on the value of a selected runtime field.
* **End While –** should be used to close the “while” loop scope.

**Extract Menu:**

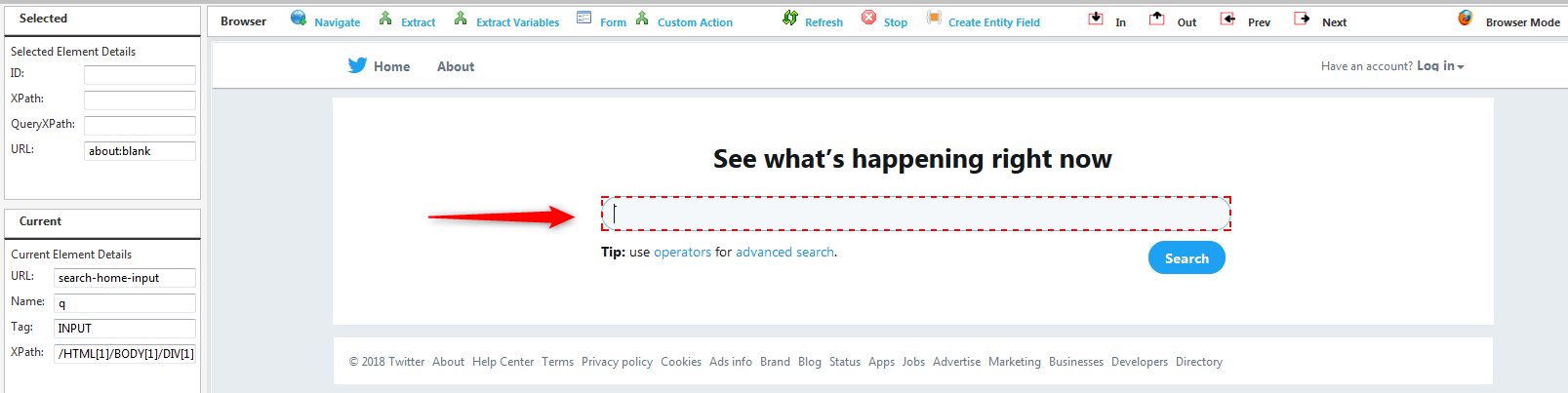
* **Save Entity** - Explicit request to save current entity’s values in the database.
* **Extract Value** – Store selected value within selected entity field.
* **Extract URL** – Extract URL from the selected HTML tag and store it in the selected field.
* **Extract Page URL** – Extract the loaded page URL and store it in the selected entity field.
* **Extract Attributes** - Extract a certain attribute from the selected HTML tag and store it in the selected entity field.
* **Download Resources** – Download various file types including images.
* **Download Video –** Download video files using specific drivers for a predefined list of web-sites.
* **Use WIC Coding System –** enter to a field an enum value (for example: object type, image type, etc.)

**Form Menu:**

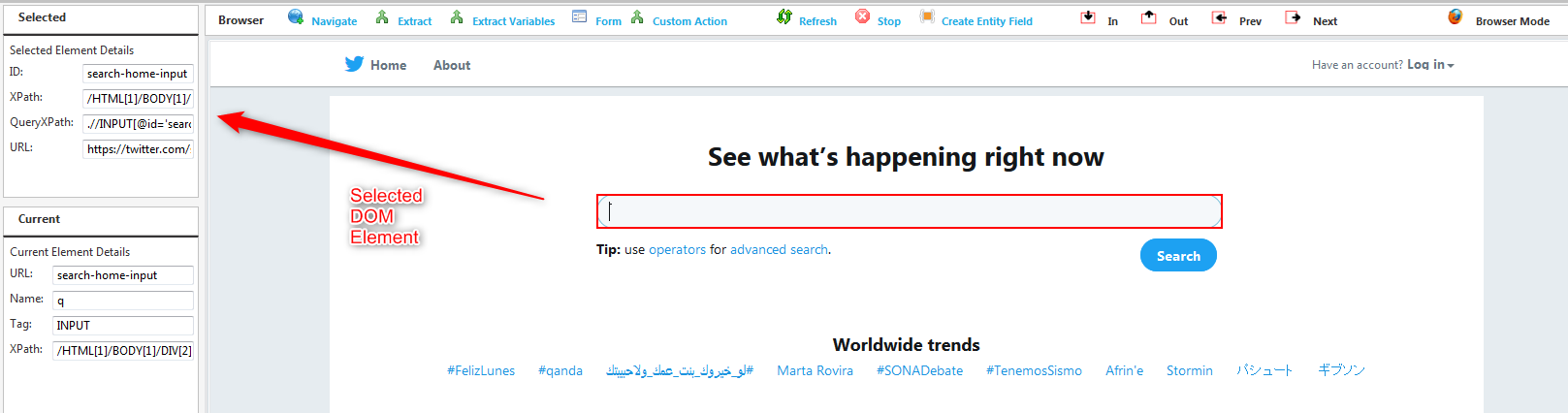
* **Enter text** – Allows entering text into selected text input box.
* **Select Option** – Allows combo-box selection.
* **Set Checkbox** – Allows marking selected checkbox.
* **Set Radio** – Allows radio button selection.

###### Creating New Steps

As the WebFlow mimic real user behavior, the different steps enable creation of actions sequence.  
Some actions are DOM based and being executed over specific HTML tags (such as “Click”, ”Enter Text”, “Extract”, etc.), other actions can work regardless to specific DOM path (such as “Load Page”, “Scroll Down”, etc.).  
In order to create DOM based steps, choose the specific element by clicking the red dotted rectangle.

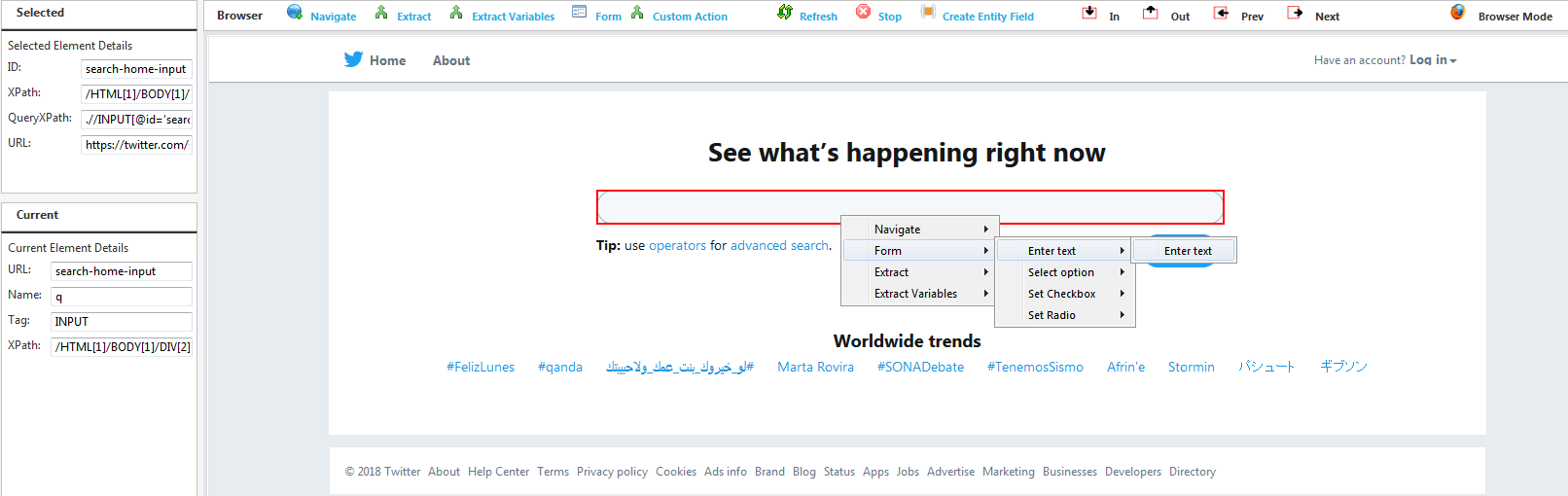


When the element is selected the red dotted rectangle’s line will become solid and the tag properties will displayed in the **Selection Panel.**



At this stage it is possible to select a specific DOM based action by:

* Right mouse click on the element.
* Click the relevant item within the browser menu.
* For example: “Enter Text” by right mouse click on the element.



* 1. Executing WebFlows

Once the WebFlow is ready, there are few prerequisites to set before it can be executed:

1. A Web-Platform should be created
2. The WebFlow should be published
3. The WebFlow should be attached to the WebPlatfrom
4. A virtual Agent should be created (optional)
5. A new collection request should be created

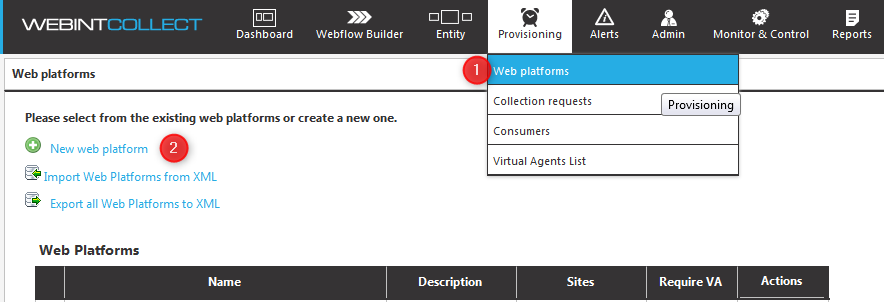
###### Create Web Platforms

Web platform structure is used to connect WebFlows of various functionality types to a specific web site. Good example for using that are Facebook, Instagram or Twitter web-platforms, that allow the user to perform different type of collections within the same web site. The user may search for posts, collect profiles’ pages, etc. while keeping the same Web Platfrom name that is presented in Webint Analytics user interface.

To create a new web platform, follow the steps below:

**Step 1**

At the upper menu click **Web Platforms** under the **Provisioning** **Menu,** thenclick **Create new CR.**

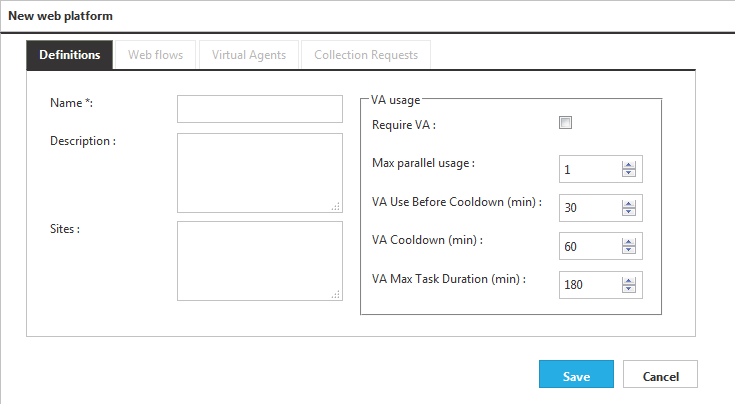


**Step 2**

In the **New Web Platfrom Editor** enter the following:

* **Name –** The web platform name, usually will be the web site name (Facebook, Twitter, etc.).
* **Description** – optional description for the web platform.
* **Sites** – optional description for the web platform’s web-sites.
* **VA Usage –** the fields within the box refer to the web platform’s virtual agent policy.
  + **Require VA** – should be checked for web platforms that contain WebFlows that require virtual agent to login.
  + **Max Parallel Usage** – indicates how many collection tasks can use the same VA in parallel.
  + **VA Use Before Cooldown –** indicates how many minutes the VA can be used before it should be offline (to avoid VA blocking by the web site).
  + **VA Cooldown –** indicates the duration of the VA cooldown in minutes, during the cooldown the VA cannot be used for collection requests execution.
  + **VA Max Task Duration** **–** indicates the maximum duration of a single task that utilize a VA. Once the task exceeds this duration, the task will be stopped.  
    This mechanism is also meant to avoid the blocking of VA’s by the web site.

Then click **Save.**



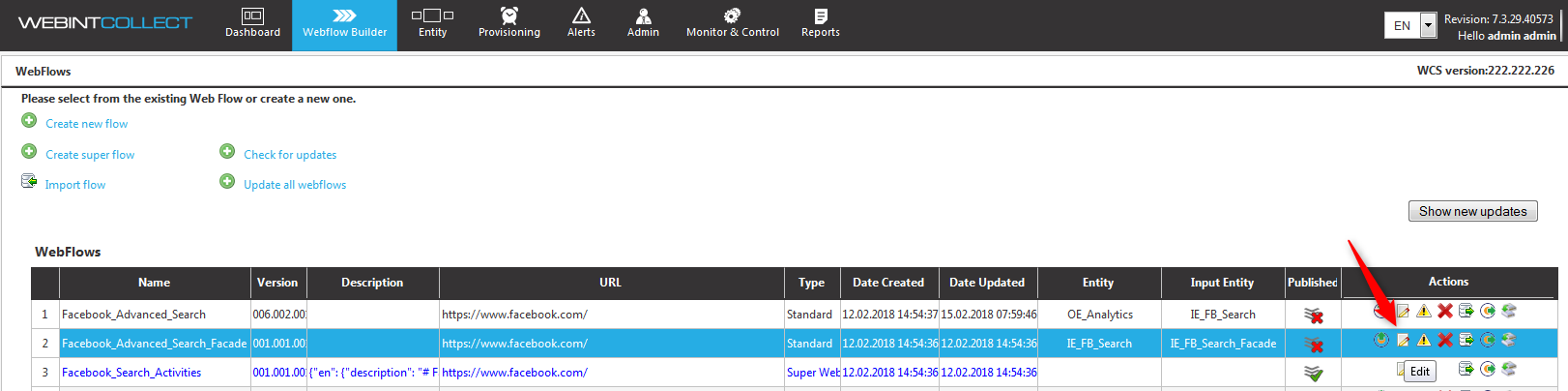
###### Publish WebFlow

WebFlow publish action meant to expose the specific WF to the web platform.

To create a new web platform, follow the steps below:

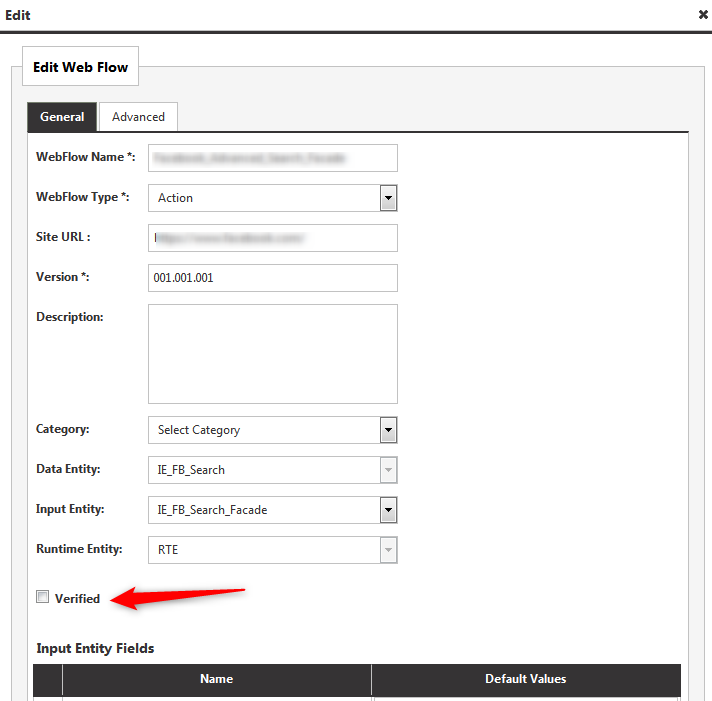
**Step 1**

At the upper menu click **WebFlow Builder**, then at the relevant WebFlow click **Edit .**



**Step 2**

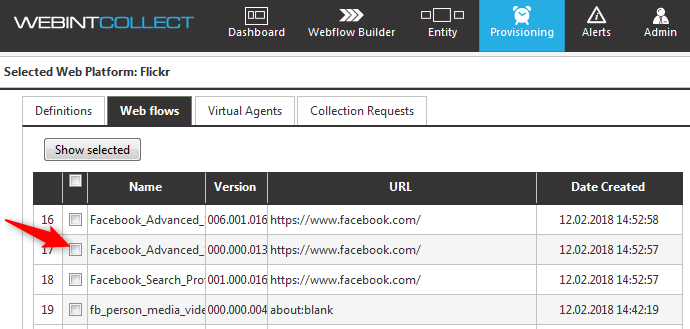
In the WebFlow editor, mark the **verified** check box and click **Save.**



###### Attach WebFlow to Web Platform

Attaching WebFlows to web platforms enable to execute the attached WF’s.

To attach the WebFlow, go to the **WebFlows tab** within the web platform, mark the relevant WebFlow’s checkbox and click **Save**.

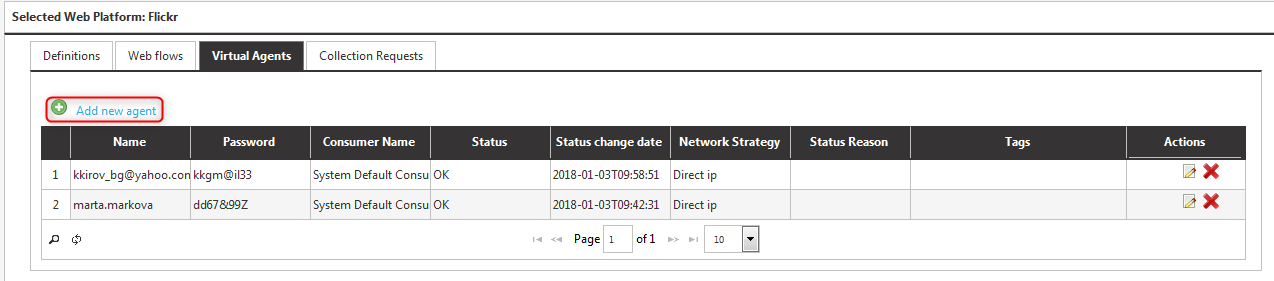


###### Create a Virtual Agent

Virtual agents can be created within a specific web platform.   
Once a virtual agent in created, every execution of a WebFlow will use the existing virtual agents (one agent per collection task) including their attached proxy definitions.

To create a new virtual agent, follow the steps below:

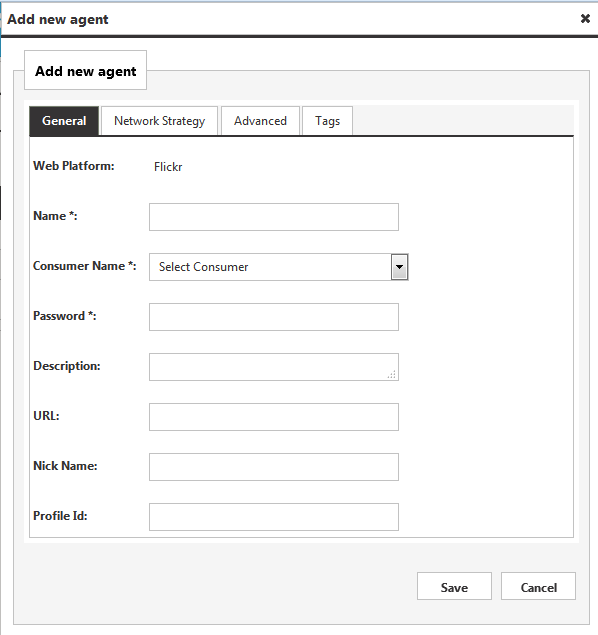
**Step 1**

Go to the **Virtual Agents tab** within the web platform, then click **Add new agent.**

**Step 2**

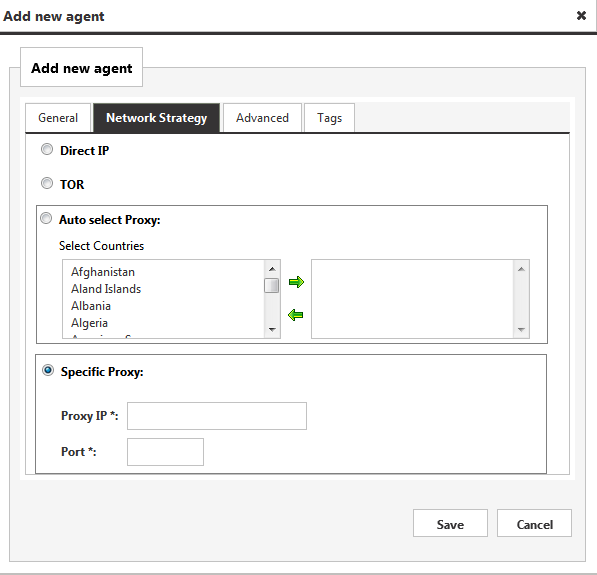
In the **New Virtual Agent Editor**, under the **General tab** enter the following:

* **Name –** The virtual agent’s exact username.
* **Consumer Name –** refers to the instance of the Webint Analytics (in case there is more than one Webint Analytics system connected). **System Default Consumer** should be selected by default.
* **Password** - The virtual agent’s exact password.
* **Description** – optional description for the virtual agent.
* **URL** – optional field for keeping the URL of the virtual agent.
* **Nick name** – optional field for keeping the nick name of the virtual agent.
* **Profile Id** – optional field for keeping the profile Id of the virtual agent.



**Step 3**

In the **New Virtual Agent Editor**, under the **Network Strategy tab** select:

* **Direct IP –** to use the IP of the used crawler unit.
* **TOR –** to use TOR and connect to Dark Web sites.
* **Specific Proxy –** to connect with a predefined and accessible proxy server. Notice that the proxy should be accessible from the system crawler units.

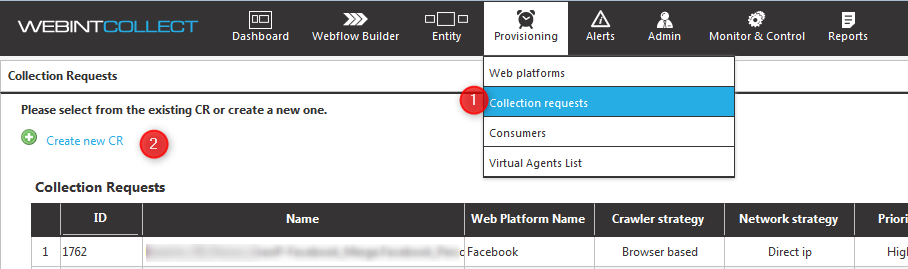
###### Create a Collection Request

A collection request is meant to allow execution of a WebFlow instance with a set of predefined input parameters.

To create a new collection request, follow the steps below:

**Step 1**

At the upper menu click **Collection Requests** under the **Provisioning** **Menu,** thenclick **Create new CR.**

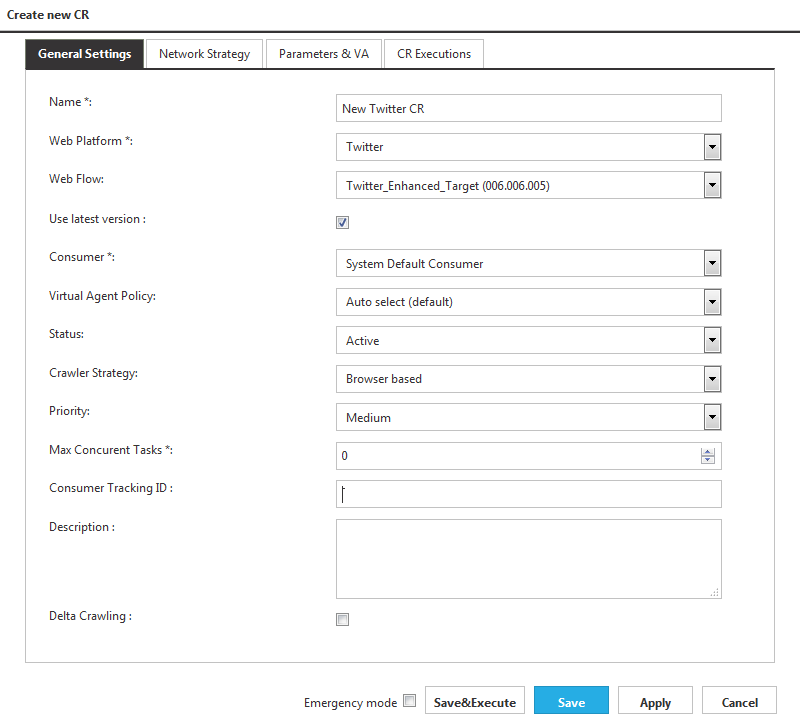


**Step 2**

In the **New CR Editor**, under the **General settings tab** enter the following:

* **Name –** The CR name.
* **Web Platform–** which web platform should be used.
* **Web Flow** – which WebFlow within the web platform should be used..
* **Use Latest Version** – should be marked in order to use the latest version of the selected WF.
* **Consumer** – refers to the instance of the Webint Analytics (in case there is more than one Webint Analytics system connected). **System Default Consumer** should be selected by default.
* **Virtual Agent Policy** – if there is no specific VA to be used, should be left with **Auto Select.**
* **Status** – should be remain as active.
* **Crawler Strategy** – should be remain as browser based.
* **Priority** – should be modified in case there are other concurrent executions that are less urgent.
* **Max Concurrent Tasks** – determine how many collection tasks (browser instances) can be opened in the same time. The default 0 meant unlimited.
* **Consumer Tracking ID –** used only for internal processes.
* **Description –** optional description of the collection request.
* **Delta Crawling – optional,** can be selected when delta crawling is implemented in the WF, but is not mandatory.

Then click **Save.**

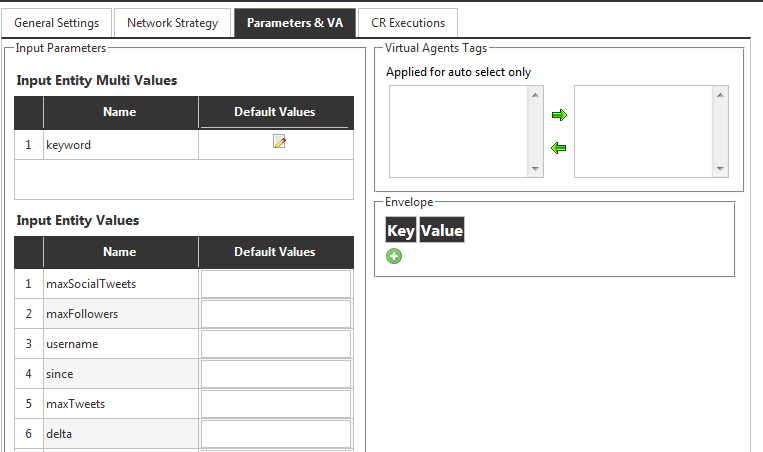


**Step 3**

In the **New CR Editor**, under the **Parameters & VA tab**:

* **Input Entity Multi Values –** add here the relevant URL’s or keywords for the relevant functionality type field.
* **Input Entity Values –** add here the value for each field.

Then click **Save.**



* 1. WebFlows’ Execution Info

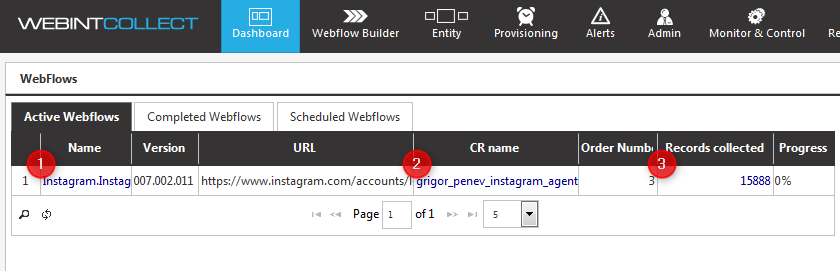
In order to monitor and view WebFlows’ execution that are currently running or completed, one should review the info that is presented in the WebFlows Widget in Webint Collect Dashboard.

**Active WebFlows**

All the running collection requests can be viewed within the Active WebFlows tab within the WebFlows widget.

This view include 3 shortcuts (as marked below):

1. Clicking the **Name** column’s value will navigate to the WebFlow within the WF builder.
2. Clicking the **CR Name** column’s value will navigate to the collection request page.
3. Clicking the **Records Collected** column’s value will navigate to the content report interface which presents the collected records in a tabular view.



**Completed WebFlows**Once the WebFlow finishes its execution, a new record related to the completed execution will appear in the Completed WebFlows tab within the WebFlows widget.

This view include 4 shortcuts (as marked below):

1. Clicking the **Name** column’s value will navigate to the WebFlow within the WF builder.
2. Clicking the **CR Name** column’s value will navigate to the collection request page.
3. Clicking the **Status** column’s value will navigate to the execution report interface.
4. Clicking the **Records Collected** column’s value will navigate to the content report interface which presents the collected records in a tabular view.

