# Prof Catalyst Maintenance

This document covers system maintenance topics.

## Paid Services

The whole system relies on several paid services. Disabling even one of them can lead to unstable system operation.

* **Mongo Atlas**  
  Main database of the system. Most important feature used is integrated full-text search engine.
* **Linux Servers**Responsible for hosting system sites: FTP, UI, and SRV (crawlers API endpoint).
* **Windows Servers or PCs**  
  Crawler instances are located here. It can be any Windows machine connected to the Internet and have installed corresponding software.
* **Dolphin Anty Browser**Main browser used by crawlers. Account and proxy management is relied on its infrastructure.
* **Proxy Servers**  
  Some social networks (like instagram) have anti-crawl policies based on IP addresses. In that cases you must use as much different IPs as possible.
* **Social Network Accounts**

Some social network data are available only for authorized users. So, you must have a list of registered accounts to perform crawling.

## Crawler Instance Configuration

Crawler application is written and tested for Windows machines. It could be either Windows server or any private PC. It only requires installation of specific apps and stable internet connection.

### Initial Configuration

1. Изображение выглядит как текст

   Автоматически созданное описаниеDownload and install Google Chrome browser: <https://www.google.com/chrome/>
2. Download and install Dolphin Anty browser: <https://anty.dolphin.ru.com/en/>
3. Run Dolphin browser and login as: [andrey@roibees.com](mailto:andrey@roibees.com)
4. Create a Crawler folder and upload CrawlerApp files into it.
5. Make a shortcut to CrawlerApp.exe file.
   1. Make sure the working directory is set to Crawler app folder.
   2. You can pass “--kill-drivers” argument to clear stuck drivers in case of failures.
6. Additional or specific configuration is in “Configs/config.json” file.
7. To start crawling make sure Dolphin is running and launch the shortcut.
8. Use Ctrl+C to store any unsaved data and terminate the crawler application gracefully. Note that the crawler needs some time to complete current tasks before termination.

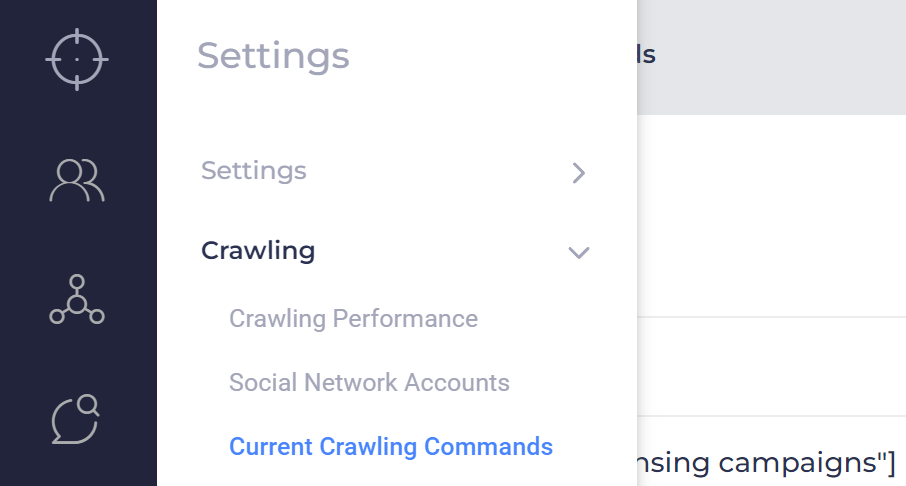
### Configuration File

All main configurations are in “Configs/config.json” file. Here is a list of possible options:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Type** | **Default Value** | **Comment** |
| CheckUpdateInterval | Int | 300 | Crawler update check interval in seconds. |
| ApiUrl | string | "https://sc-api-srv.profcatalyst.com/api" | URL to crawlers API endpoint. |
| ApiMail | string | "qa@roibees.com" | Endpoint login. |
| ApiPass | string |  | Endpoint password. |
| PrioritizedTaskCount | int | 50 | If a Command has fewer Tasks, try to finish it completely. Otherwise use Task priorities. 0 to disable. |
| Threads | int | 1 | Simultaneous crawler threads count. |
| WaitTimeout | int | 300 | The maximum time in seconds to wait for browser operations before fail. |
| RetryTimeout | int | 150 | Time in seconds until the next attempt to execute the failed or delayed task. |
| RetryAttempts | int | 10 | Maximum number of attempts before skipping a failed task. |
| MultipartSizeThreshold | int | 1000000 | Threshold in bytes to accumulate crawled data before sending to the server. |
| MultipartFilesThreshold | int | 20 | Maximum number of image files to accumulate before sending to the server. |
| MultipartVarsThreshold | int | 1000 | Maximum number of data objects to accumulate before sending to the server. |
| StorageApiThreads | int | 1 | Simultaneous storage threads count. In most cases 1 is enough. |
| StorageApiSendInterval | int | 1 | Interval in seconds between storage requests. |
| Strategy | string | "Default" | Next Command selection strategy:   * Default – Take commands sequentially and try to execute completely. * Urgent – Take the newest commands and leave the previous ones unfinished. |

### Crawlers Monitoring

Each crawler instance generates unique ID on start with template: country, colon, IP-address, colon, 4 random bytes. Example: “US : 144.202.125.7 : 05D78D14”.

To view lists of Commands and check those reserved by crawlers, open the Settings page, and navigate to Current Crawling Commands section.

On that page you will see every available Command, reservations, and progress of execution.

Crawler instances report progress every minute. If the crawler stops reporting for more than 5 minutes, the monitoring system sends an email notification.

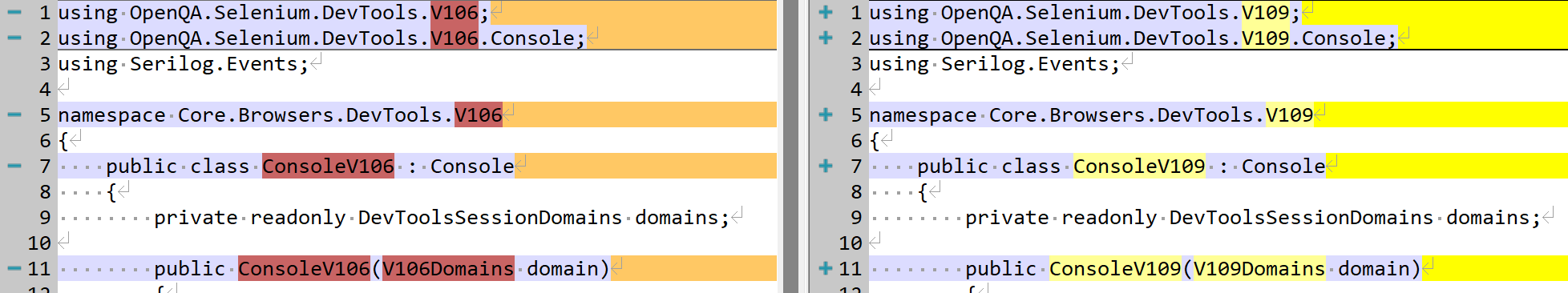
### Development and Upgrades

The crawler depends on applications that are constantly updated. Most often, these are browser versions that are updated on their own. In new versions of Selenium, the protocol may be slightly changed. So, you should monitor for updates in the Selenium.WebDriver NuGet package:

Изображение выглядит как текст

Автоматически созданное описание

If there is a new version of the driver, then the package must be updated in all projects to the same version. Usually, in a new version of a package, the oldest version of the protocol becomes obsolete, and the latest version appears. You should make corresponding changes to version specific classes:



After, code could be compiled, tested, and deployed.

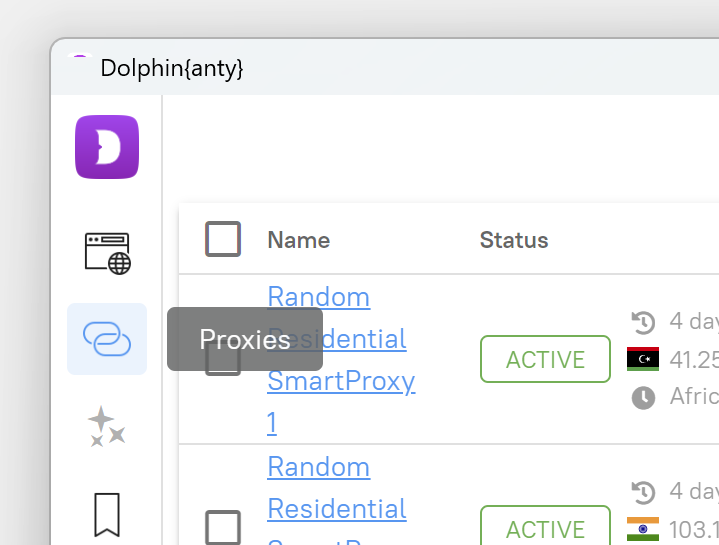
The crawler implements custom continuous delivery protocol. There is a Deploy application project that packs new Crawler into zip-archive and upload to API-endpoint's storage. All connected crawlers poll the endpoint and download and apply update automatically.

## Proxy Configuration

All proxy configurations are done in the Dolphin Anty browser. Proxies are set independently of accounts (or browser profiles in Dolphin terms), but each account can be associated with at most one proxy. Dolphin supports http, socks4, socks5 and ssh proxy protocols.

* For anonymous accounts, it is recommended to use a proxy with the Rotating Session type to randomize IP addresses.
* For authorized accounts, it is better to use Sticky Session type to minimize IP changes and possible account bans.

To add a proxy to Dolphin, you need to go to the Proxy section by clicking the corresponding icon in the left menu:



Next, you need to click on the “+” button in the upper right corner, next to the search bar. After clicking, you will see a dialog box for adding a proxy.

Изображение выглядит как текст

Автоматически созданное описание

Enter the address in the appropriate field in the add proxy dialog box and click the “Add” button in the upper right corner. All possible addition formats to the left of the field are displayed.

After adding, our proxies will appear in the general list of the Proxy section, and they can be used when creating accounts.

It is recommended that you provide a friendly name for the proxy, as it is much easier to identify a proxy by its name. You can add it in the proxy editing dialog box, where you can also check their status (operability). To do this, simply click on their current name, which is underlined, and an edit dialog box will appear in front of you.

## Social Network Account

Account management system is relied on Dolphin Anty browser profiles infrastructure. To avoid collisions, it's better to use the rule of thumb: create separate browser profiles for each social network account.

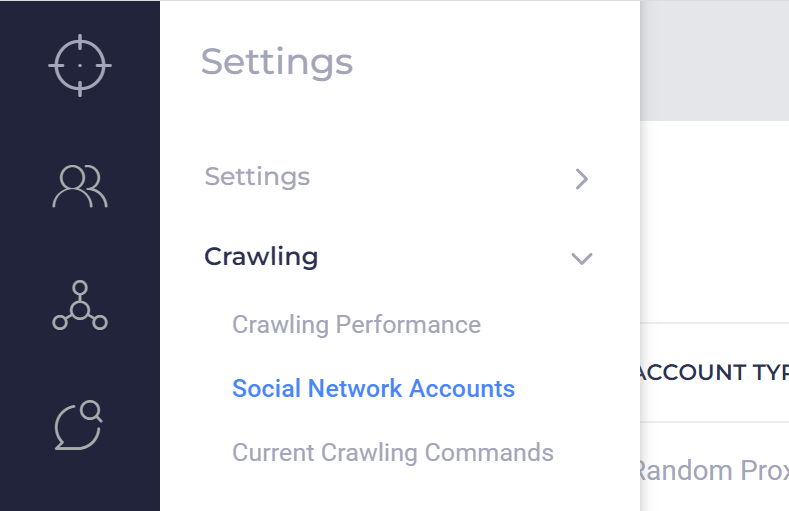
The best practice is to follow the following rules:

1. Set the browser profile name using the template:
   1. short social network id, space, account's user id, space, browser profile id.   
      Examples: “fb 100087077737619 32758482”, “ig blufovmeoplas1975 25452615”.
   2. For anonymous accounts (browser profiles not authorized in any social network) template: anonymous word, space, browser profile id.  
      Example: “anonymous 27634866”.

Browser profile id can only be obtained after creation, so you must edit the profile to be able to copy it. Naming system makes it easy to navigate through accounts.

1. Leave the status field empty. Here is the recommended usage of statuses:
   1. No Status – browser profile configuration is incomplete,
   2. New – browser profile is configured, and account is tested, but profile is not registered in crawler system,
   3. Ready – browser profile is registered and already in work,
   4. Ban – the account is blocked or logged out in some way.
2. If necessary, select proxy from a predefined list.
   1. Note to use random or rotating proxies for anonymous accounts, and constant or sticky proxies for authorized accounts.
   2. To reduce traffic usage, run the browser profile after creation and disable images in Chrome settings.
3. Paste cookies json-data if you have it. Otherwise, it's better to run the profile after creation and try to log in manually. After successful login, Dolphin browser will store necessary cookies automatically.
4. Set the language property to manual, English.

After creating a browser profile, you need to register it in the Crawling system. To do it open Settings page and open Social Network Accounts section:

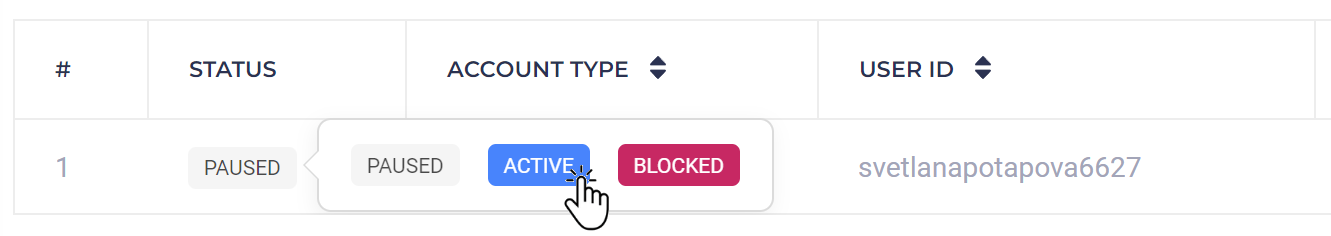


Изображение выглядит как текст

Автоматически созданное описаниеClick "Add Account" button and fill the form:

* Account Type – desired social network or Random Proxy for anonymous accounts.
* Dolphin Anty browser profile ID – profile that will be started to perform crawling task.
* Other fields are optional:
  + User ID – account’s id in social network terms. Used to faster auth checks and filled automatically if omitted.
  + Login/Password – account credentials to try to perform automatic authorization if necessary.

To get your account up and running change its status to Active:



Crawler instances request Active account and use it. If it determines account as not authorized or banned, account automatically gets Blocked status and excludes from usage. You can still resolve the blocking issue manually and set status back to Active.