Workstream 2

Monitoring online shops

Final report

Department of Addictology, First Faculty of Medicine,
Charles University in Prague and General Teaching Hospital in Prague

Co-funded by
the European Union

Coordinator:
Daniela Kmetonyová (CUNI)
Martin Pažitný (CUNI)

Date: 30/09/2015

Table of contents

Search Engine Helper and Site Scraper Software

Abstract - This report is brief description of newly developed software under the project I-Trend Workstream 2 Monitoring of online shops. The software consists of two modules - search engine helper and site scraper software. The software is used to search for keywords based on automatic search and further categorization of found websites and for gathering content from selected web sites and saving them into database scheme.

The software has been written in python programming language [Rossum] using few non-standard third-party software libraries (Django, Scrapy, Webkit Scraper, Dryscrape, Scrapely etc.). The user frontend of the software is database web application. A Django [Django] framework is used for the frontend and for database access. Database system can be any relational database with python bindings. There are several background automatic processes e.g. analytics gathering, searches, categorization, scraping and screenshot taking. All of these processes access the database system via Django.

There are two main user frontend modules: search engine helper and site scraper. The search engine helper provides user friendly interface to various search engines (google, yahoo, etc.) and stores results in the database. The results can be later accessed, categorized or exported by the user. The site scraper is a tool for extraction of an useful user-defined information from selected websites (the results of the search engine helper software).

Both modules use the same mechanism for accessing pages on the world wide web. Web pages are usually mixture of technologies including: protocol HTTP, HTML, CSS, JavaScript, etc. In order to properly render their content, web browser components are needed. As the rendering core a Webkit Scraper is used, which is based on Webkit engine (formerly developed by Apple for the Safari web browser).

In *the search engine helper* an user interaction with search engines is mimicked and the search results are gathered and stored into the database scheme. Content of each newly encountered result is compared against list of keywords and then enlisted in category based on keywords. In selected categories more information about the result is gathered: analytics, screenshot, domain registrar information, location of the web server. There is an option to

change category of each result by the user. The efficiency of keyword based categorization ought to be evaluated and categorization algorithm improvement based on more comprehensive statistical inference should be proposed.

The site scraper interacts with the results enlisted in particular category – web shop URL, with aim to gather all of the accessible web pages under each URL listed. This is basically web crawler behavior. Whole process is governed by the third party software library Scrapy, which has been integrated into our software. Each gathered page is then processed using instance-based learning algorithm [Yanhong and Liu; Muslea et al.], implemented in the Scrapely software library. This algorithm allows user to match the visible text on the sample web page with database attribute. Algorithm then tries to gather annotated data on other pages within the same site based on annotated sample site. The gathered data is then stored in database scheme. There are several limitations while trying to gather structured data from loosely-structured pages. The whole algorithm ought to be improved in order to provide more reliable data.

References

[Rossum] G. van Rossum, Python reference manual, Technical Report, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

[Django] Django core team, Django: A Web framework for the Python programming language, Django Software Foundation, Lawrence, Kansas, U.S.A., http://www.djangoproject.com, 2011

[Yanhong and Liu] Yanhong Zhai , Bing Liu, Extracting Web Data Using Instance-Based Learning, World Wide Web, v.10 n.2, p.113-132, June 2007

[Muslea et al.] Ion Muslea , Steve Minton , Craig Knoblock, A hierarchical approach to wrapper induction, Proceedings of the third annual conference on Autonomous Agents, p.190-197, Seattle, Washington, United States, April 1999

Appendix A: Installation instructions

This manual is valid for Debian or Ubuntu server operating systems.

1. Installation of system wide dependencies:

sudo aptitude install python-pyside.qtwebkit python-virtualenv ipython python-lxml python-crypto python-unidecode xvfb git python-psycopg2

OFDT environment : sudo *apt-get* install python-pyside.qtwebkit python-virtualenv ipython python-lxml python-crypto python-unidecode xvfb git python-psycopg2

2. Intialization of the environment:

```
git clone https://pborky@bitbucket.org/pborky/scrap.git
cd scrap
make initenv
bin/activate
make dev-setup
```

You will be asked for administrator user name and password.
Now you are ready to start the server:

```
sudo -u www-data xvfb-run -a bin/python manage.py runserver 0.0.0.0:8000
```

Home NPS Publications Survey About us

Methodology

Department of Addictology, First Faculty of Medicine,

Charles University in Prague and General Teaching Hospital in Prague

Co-funded by the European Union

Daniela Kmetonyová

Martin Pažitný

Kateřina Grohmannová

Vendula Běláčková

Viktor Mravčík

Tomáš Zábranský

1. Objectives

The general objective is

(i) to analyze the availability of synthetic psychoactive substances through the specialized internet websites – web shops market places available on surface web (not Deep web the part of the Web not indexed by search engines[1]) and

(ii) To identify the most marketed substances in each member country.

(iii) to provide a better estimation of market size

The methods of data collection will consist of monitoring online retail and wholesale shops with NPS available in national languages (or other relevant languages for the partner's country), their products and, to the extent possible pricing and marketing strategies.

The analysis will build upon the data previously collected by EMCDDA via its snapshot survey (http://www.emcdda.europa.eu/publications/scientific-studies/2011/snapshot) and will update its methodology with recent developments in search technologies and with findings from WS FORUM and WS SURVEY in terms of identified market triggers. The results will be interlinked with the information collected via WS FORUM and WS SURVEY on demand for / use of new psychoactive substances. Information collected will feed into final Top 10 and technical folders on selected compounds.

1.1. First objective: MONITORING new psychoactive compounds retailed through the internet websites

Monitoring is based on quantitative and qualitative indicators.

Expected outputs are the following:

- Number of online shops offering NPS relevant for customers in each partner country (covered by general monitoring including mostly TOP 10 NPS (commercial names)+relevant names of NPS from online fora + other relevant names for each country for some other reason).
- Report on the most marketed substances online in partner countries
- Collection of supply side data to feed into the technical folders on chemical compounds available on the internet (see workstream Top 10 and Technical folders).
- Try to identified and remove duplicities in market size and estimate the true numbers (e.g. by finding web shops operated by one owner)

[1] https://en.wikipedia.org/wiki/Deep_Web_%28search_indexing%29

1.2. Second objective: qualitative and quantitative ANALYSIS of marketing strategies of the retailers and the wholesalers on the internet

- To provide an overview of marketing strategies to the extent possible and with special attention to pricing of particular products
- To provide an overview of marketing strategies targeting specific age groups of customers and the age (and other type of) restrictions the retailers (self) impose
- To categorize the shops based on their marketing strategies
- To describe the most common marketing strategies of the NPS retailers in terms of product labeling and purchasing forms (bath salts, fertilizers etc.), description or classification of compounds and effects
- To interlink the available supply data on marketing strategies with the data obtained in WS Forum and WS Online survey and analyze the driving forces of the market and relationships between the following variables:
    - marketing strategies for substitute/complementary products
    - labeling strategies and the demand for the substance

1.3. Third objective: to enrich the standardized METHODOLOGY of data collection with new techniques and tools, especially software

- provide analytical reasoning for individual items collected within the project and hereby to enrich the data within standard EMCDDA monitoring system
- propose optimization techniques for ongoing / continuous and consistent monitoring that minimizes the input of human resources and maximizes automation of the processes
- clear data from duplicate results shops

2. Workstream activities

2.1. Defining the scope of observation, validation of a common methodology, distribution of activities to each partner in this workstream

2.1.1 Defining the scope of observation

2.1.1.1 Agreement on Research Questions covering all 3 WS objectives

a) Questions related to objective no. 1 – MONITORING

Reports on most marketed substances

By "most marketed substance" we mean products offered by highest number of selected most popular online shops targeting partner country. If several substances are offered by same number of online shops, then the more marketed substance will be the one, which was offered by more shops in previous monitoring period.

- What are the top 10 most marketed substances in each of the project partner countries?
    - How do the top 10 substances in partner countries change over the time
    - The Czech Republic
    - France
    - Poland
    - The Netherlands
    - The United Kingdom

Data to feed into technical folders

- How many shops offered the substances in the observed period?
- Under what marketing brands have the substances been sold in these shops?
- In what quantity and price ranges have the particular substances been sold?

b) Research questions related to objective no. 2 – MARKETING STRATEGIES

Questions related to shops

- What are the most popular online shops offering NPS relevant to project partner countries?
- What is the webpage shop ranking in selected popularity measures?

Questions related to price

- ⊃ What is the variability in price / quantity ratio of selected substances for partner countries (highest and lowest price, mean)?
- ⊃ What is the relationship between price and other online marketing tools (for list of tools see methodology of analysis of marketing strategies)?

Questions related to substance labeling

- ⊃ What are the most common strategies for product presentation in partner countries (salts, chemicals, collector items et al.)?
- ⊃ What is the variation in product labeling for particular chemical substances for partner countries?
- ⊃ What is the variation of description of effects for particular chemical substances for partner countries?
- ⊃ What is the variation in advertising for partner countries?
    - marketing the substance as “new”
    - marketing a “price reduction”
    - offering “extras”, such as additional product for free et al.
    - "replacing strategy" - offering the product as substitute for previously banned substance (e.g. alternative mephedrone instead of banned mephedrone)

Questions related to legislative responses

- PRODUCT: advertising, "replacement strategy"
- SHOP: restriction on shipments, location of the shop, warnings on legal issues

Questions related to harm reduction strategies

- ⊃ What percentage of shops in partner countries applies age restrictions on sales?

2.1.1.2 Agree on data collected and database structure

- ⊃ By data collection we should be able to answer our Research Questions
- ⊃ For list of all collected data please see chapter 2.2 + file WS SHOPS_database_structure.xls
- ⊃ Each collected data will be coded, so it’s easier to run statistics
- ⊃ Getting of some information may not be possible to automate, in this case human interaction will be needed

2.1.2 Validation of common methodology (monitoring + analysis of marketing strategies)

Current EMCDDA searching methodology:

Currently, the EMCDDA snapshot survey consists of manual search of selected search engines by selected key words (search strings). Then the results of the search are checked and if a web shop selling NPS is found, further details are monitored and filled into xls file with pre-defined structure.

Period: usually once a year

Search engines used:

- Local google search engine (for Czech republic – google.cz)
- Bing.com (previously also metacrawler.com, yahoo.com)
- Local most popular search engine (for Czech republic – seznam.cz)

Key words: identification of key words and developing search strings according to national context; the search strings should identify 1) sites using common generic terms to refer to new drugs, 2) sites selling smokable herbal or resin-type products that may contain synthetic cannabinoids, 3) sites selling specific substances: kratom, salvia divinorum, mephedrone (4-MMC), methylenedioxypyrovalerone (MDPV), gamma hydroxybutyrate (GHB) and gamma butyrolactone (GBL), and hallucinogenic mushrooms.

Process of the search:

Search first 100 hits found,

- If found another shop in next 20 hits, search another 20 hits
- If not… END

Database: xls file (same for all countries)

Only shops with local language interface

New methodology

The methodology is to be updated in terms of (i) items collected, (ii) automation of the process – new software tool. The methods used will as well shape the content of the database and the scope of observation. In summary this means that the search (scraping) will be conducted on two levels: the level of shops using global search via search engines, the second level of products, out of which three databases will be an outcome. The product analysis will be conducted for selected shops only.

⊃ **Monitoring of the web shops via search engines**

New software tool will be developed for automatic monitoring of web shops selling new synthetic drugs (NSD) found by search engines. The software will conduct automatic search of specific keywords combinations in selected search engines and will return the results from all search engines in one place (online platform). Also several categories will be created within this online platform for easier categorization and interpretation of the identified web shops.

At first all the results will be moved to one of the following categories:

1/ E-shop –all the results will be checked against a "definition of web shop" and if it fits, the web page will be added to category. By "definition of web shops" we mean group of words which usually a standard web shops contains, e.g. Price, Basket, Check-out etc. This is the way for the software to know that the found web page is a most likely "web shop". Here we assume that thanks to this filter we can identify and further monitor new web shops selling NPS more quickly.

2/ Other – all the other results

Then from E-shop category on regular basis the results will be manually moved to several other categories that will be created within this online platform for categorization of relevant web shops with the aim to minimize duplicities and thus getting better understanding of real market size (category Other can be checked irregularly).

Selection of relevant web shops for further monitoring:

First and most important criterion of inclusion is interface in national language. If this criterion is missing, than it is to possible use additional criteria such as that e.g. the web shop has strong link to the partner country. So e.g. partners country domain name is used (e.g. .nl) or the shop is apparently based in the country (by address in contact information) or people speak about it on forum and so on. In such cases the interface in different than national language is acceptable.

Distribution and characteristics of web shops in different categories:

- Approved category: consist of web shops which have unique design (no other identified web shop has the same design) and are properly working
- Duplex URL: This category consist of web shops that have their equivalent in category "Approved e-shops", but the URL format differs only with https instead of http, for example https://rcpoland.eu/ vs http://rcpoland.eu/ or URL address is missing another characteristic. These web shops are identical, except different http protocol. In "Approved" category are kept web shop with "https" protocol (more technologically advanced), other variations are kept in category „Duplex URL".
- Dead E-shops: in this category the web shops which were previously active and stopped working are kept. After entering the URL, you can usually see message: "Page is not available" or variation of this message.

- Redirect: In to this category the web shops which are automatically redirecting to another web shop/URL are kept. For example: http://rc.cx/ is automatically redirected to http://research-chemicals.pl/?wiekok=yes.
- Parallel e-shops: This category includes web shops, which have the same web page design and content as other web shop(s) in „Approved“ category, but have different URL address. Variation may be for example in a domain name: kickass-stuff.net vs kickass-stuff.org or the URL address may be completely different, for example http://www.3mmc.me/ vs http://3momc.eu/. But what is important is that the design and content of web shops in both cases is identical. In „Approved“ category just the web shops with unique design and content are kept and others shops with identical design and content than the one in „Approved“ category are moved to „Parallel“ category. The more popular web shop from group of these shops is moved to category „Approved“ and the rest is moved to category „Parallel e-shops“. For popularity estimation we use Alexa Global ranking. Shops which are hosting in public market places such as 'otwarte.pl' usually stay in approved category. Market places provide shop templates and this is why shops on public market places have similar design. Only identical shops from those markets are moved to “Parallel” category.

New software will periodically collect print screens of home page from all relevant web shops. Those will be then used to help with the categorization described above - as source for tracking and describing of web shops' design changes and also assessing web shops lifetime.

Several other characteristics about the relevant web shops will be also automatically gathered by the software, e.g. popularity characteristics, information on IP location etc.

Period: weekly

Search engines used: most popular search engines defined by each partner country

Most popular search engine is the one visited by most of the people – for selection criteria please check following traffic and/or web ranking pages:

- http://websitetrafficspy.com
- http://www.semrush.com
- http://www.pageranking.org
- http://www.checkpagerank.net
- http://www.prchecker.info
- http://www.alexa.com

Key words: List of key words for the search will be agreed by all partners (National Top 10).

Partners will be asked to provide translation into their native languages. If technically possible, more than 10 keywords can be used for search (starting with Top 10 and gradually adding more keywords over the time, which are relevant for the partner’s country).

Process of the search:

Search the first 100 hits - here we suggest covering only the first 100 hits because certain search engines limits the number of searches and the payment is needed for showing more results, mostly when conducting many searches over short period of time (e.g. google). Also this limit seems reasonable in terms of server usage.

Database: will be agreed (same for all countries)

⊃ **Monitoring the selected web shops and their products**

Selection procedure: Each partner will select 6 web shops selling NSD based on the criteria below; subsequently, an automatic script (software) will be developed by CUNI that would scrape the required information from each web shop.

The selection of those web shops should be based on the following criteria:

- Web shops were found by standard EMCDDA monitoring and stay active until today (check against EMCDDA monitoring data).

- Web shop is more popular than the others among internet users for partner country.

For popularity comparison of web shops please follows instructions below:

1- Primarily compare the characteristic GLOBAL_RANK (alexa.com) for the 2nd name domains (e.g. www.domain.com).

In case there is a 3rd name domain (e.g. sub.domain.com) in your list, the characteristic GLOBAL_RANK gives you only the popularity of webhosting 2nd name domain.

To be able to compare 3rd name domain with other web shops, please go manually to www.alexa.com, enter the 2nd name domain and look to the "**Where do visitors go on ....(3nd name domain)...?** " section (later we will try implement this in our software for automation).

In case you are able to find your 3rd name subdomain, take the % assign to it and divide GLOBAL_RANK with this % - this is how you get the popularity of 3rd name domain. (e.g. GLOBAL_RANK = 10, % of your subdomain is 50%, total GLOBAL_RANK of your subdomain is 10/0,5 = 20)

In case you can't find your 3rd name domain name in this section, we assume that the site is not enough popular (not enough traffic), so it is not relevant to us.

2- In case you can't choose between the shops, please use characteristics MONTHLY_USERS a MONTHLY_PAGEVIEWS from http://netai.net.websitetrafficspy.com/.

3 - In case none of the above mentioned characteristics are available for the web shop, please use characteristics DAILY_PAGEVIEW from http://www.websiteoutlook.com. You can also find here the data for 3rd name domains in section "Top Subdomain" (if your domain name is not available, we assume again, that it is not popular enough).

4 - In case none of the characteristics are available for the web shop, we assume, that the web site is not enough popular and that is the reason none of the web sites monitoring internet traffic have measured it.

(There is also one optional suggestion to use http://www.semrush.com/, but this doesn't work e.g. for Czech shops at all, so we won't use it for Czech republic, but it may work for other countries).

Also the final selection should be consulted with WS FORA regarding the popularity of the web shops (if the data are available).

The selection shall respect the ratio of "marketed/commercial web shops" and "RC shops" in the list of e-shops.

- "Marketed / commercial web shop" is a site where branded product are numerous and more indicated than in a RC shop. It is selling NPS mostly under commercial names as "Funky", "Cocolino" and the chemical compound of the substances are mostly unknown to the customer. Often a commercial web shop is a "smart shop" or "headshop". They sell a lot of stuff around marijuana including grinder, tee-shirt and so on, lot of branded product and few chemical. "RC shop", where "RC" stands for "research chemicals" is selling NPS under their chemical names, often has the chemical names directly on the front page.

How to calculate the ratio of "marketed/commercial web shops" and "RC shops"?

- All web shops selling NPS should be divided into 2 categories "marketed/commercial web shops" and "RC shop". Then according to the ratio of those two categories, the selection should be done (e.g. we have 12 shops in total, from which 8 are from category "marketed/commercial web shops" and 4 others are from category "RC shop". The ratio is 8/4 = 4/2, thus the selection of 6 most popular web shops will include 4 most popular shops from category "marketed/commercial web shops" and 2 web shops from category "RC shop".

Period: weekly

Key words: Not exactly key words are needed here – in fact, each web shop has to be coded separately, in order for the scraping tool to know where to search for product title, price, quantity etc.

Process of the monitoring: Report of selected data on each web shop

Database: will be agreed (same for all countries)

Below is a summary table describing the difference between standard EMCDDA methodology and our NEW methodology:

Characteristics	EMCDDA standard methodology	NEW methodology
Period of data collection	1 x year	Weekly
Search engine used	o Local google search engine (for Czech republic – google.cz) o Bing.com (previously also metacrawler.com, yahoo.com) o Local most popular search engine	o Local google search engine (CZ – google.cz) o Metacrawler.com o Local most popular search engine o Bing.com o Duckduckgo.com o Yahoo.com
Keywords	Defined by each country in local language	Defined by country TOP 10 - translated into local language + other relevant keywords
Process of search	Search first 100 hits found, o If found another shop in next 20 hits, search another 20 hits o If not… END	Search first 100 hits
Database	XLS files with structure defined by EMCDDA	3 databases – agreed type of file - structure defined and agreed by all partners
Inclusion criteria	Only web shops with local language interface	Same as EMCDDA + other relevant web shops for partner country
Scope OF observation	Each found web page is monitored in detail - information about web shops and products	Search is conducted on 2 levels: - Global search by search engines will deliver list of all web shops selling NSD and their basic characteristics - Selected web shops (approved by partners) will be monitored in more details o Web shops characteristics o Product information

In case some information won't be able to get by software tool, human interaction will be needed. After programming of software tool it will be clear which information have to be monitored by human.

- **Analysis of marketing strategies methodology**

Analysis of marketing strategies will be done manually, but data gained from automatic monitoring done by the software will be also used.

Selection of web shops for further marketing analysis – data to feed into National technical folders. For the marketing analysis please chose 15% of the most popular web shops from national's list. Please remove web shops which didn't provide information about product

prices and those offering just bulk/wholesale order. Also remove web shops which are hosted on public marketspaces such as otwarte24.pl. Such marketplaces usually provide several templates for easy creation of web shop and thus the shops operating under such marketplace have often similar design. In addition add another 6 web shops, previously selected for the purchase of substances. The final number of web shops used for analysis might variate from country to country. For product analysis within marketing analysis, use only web shops where the product is available.

The methodology consists of 5 steps:

1. WEB PAGE – collecting information about selected web page answering following questions
    a. Domain owner (data collected from whois.com):
        i. Who is the owner of the domain? Is it local owner (e.g. from Czech), or foreigner?
        ii. For how long has been the domain in ownership of the current owner?
    b. Language
        i. What is the main language interface of the web page?
        ii. Does the web page support any other languages? If yes, which ones?
        iii. Is the translation correct, or is it more “google translate” with grammatical errors
    c. Design
        i. Does the design of the web page change over the time (new graphics, new features on the web page)?
    d. Legislative questions – Does the web shops contains any warnings/disclaimers regarding legal issues?
    e. Restrictions on deliveries – are there any restrictions?
    f. Interaction with customers, such as product reviews or other public communication with the vendor.
2. PRODUCT OFFER
    a. Main / Special offer products
        i. Which products?
        ii. Is there any offer of a gift when the product is purchased?
        iii. In which form is a product sold?

b. Description of the product

i. Quality of the description

- Origin (made in …)
- History (including history of use)
- Product effects
- Description of dosage

3. PAYMENT METHODS – what kinds of payment methods are supported?

a. Local bank transfer

b. International bank transfer

c. Payment by credit card

d. PayPal

e. Anonymous currency – bitcoin etc.

f. Sending money in an envelope

g. Other

4. WEB PAGE POPULARITY

i. What are the statistics from selected ranking pages?

- http://websitetrafficspy.com
- http://www.semrush.com
- http://www.pageranking.org
- http://www.checkpagerank.net
- http://www.prchecker.info
- http://www.alexa.com

5. SOCIAL NETWORKS

a. Does the web shop have its profile on the selected social network? – facebook.com, twitter.com, google+, local social network (for CZ – lide.cz?)

2.1.3 Distribution of activities to each partner

Each partner:

- ⊃ Has to agree on Research Question of this WS described in chapter 2.1.1.1
- ⊃ Has to agree on final Database Structure
- ⊃ Has to agree on NEW methodology for monitoring web shops and methodology for analysis of marketing strategies
- ⊃ Has to deliver translation of keywords describing NSD (Top 10) which will be used for searching web shops via search engines
- ⊃ Has to select 6 web shops for which a new software for detail monitoring will be created
- ⊃ Has to deliver key words describing "definition of web shop" and their translation
- ⊃ Each partner will check the database for possible inconsistencies in the native languages (example: if a particular webpage is not coded appropriately, what might be the problem in key words?); the monitoring will require regular checks from the side of the partners, in order to retrieve quality data, and correct the software tool.

2.2. Analysis of possible enrichment of standardized tool and techniques for monitoring of internet shops

In contrast to the old EMCDDA methodology our NEW methodology brings to the process of monitoring of online shops and the supply side data the following enrichments:

- Continuous monitoring of the market by the software tool (weekly update)
  - Instant access to the data
  - Easy categorization of the results
  - Details on different shop characteristics (popularity, whois information etc.)
  - Print screens of the home page of selected shops
  - Scraping product data from selected web shops
- Enrichment of data collected
  - The EMCDDA Snapshot survey contains the following variables:
    - Date of search
    - Search engine used
    - Search strings used

- URL
- Name of shop
- Does the shop ship to EU?
- Is there restriction on deliveries? & which
- Other comments about the shop
- Apparent country
- Supported interface languages
- Disclaimer / warnings on legal issues & which
- Product info / product name & chemical name

- However, in order to answer the research questions, expansion to the following variables is suggested by CUNI to the final list of all variables (see the XLS file for details):

SEARCH ENGINES

- E1: Date of search
- E2: Search engine used
- E3: Search string used (approved keyword by partners)
- E4: URL of identified web site
- E5: Sequence number in which URL was listed in the search engine results
- E6: IP address of the domain (home page of the selected website)E7: IP address country code
- E8 – E9: Who is the domain owner?
- EE10: From when is the domain registered by current owner?
- E12:E17: Web page popularity ranking according to selected measures

SHOP

- S1: Date of search
- S2: Search engine used
- S3: Search string used (approved keyword by partners)
- S4: URL of identified web site
- S5: Name of the web shop (stated on the web page)
- S6: Does the shop ship to EU countries?
- S7: Does the shop have any restrictions on deliveries?
- S8: Which restrictions?

- S9-S11: Contact information of web shop (if available)
- S12: Country code according to S11
- S13: Supported language interfaces
- S14: Is the web site correctly translated? Or is it more “google translation”?
- S15: Does the web contain disclaimer/warning on legal issues (also including e.g. information that the products which are sold are not for human consumption and that the shop is not responsible in case the product is used different way) etc.?
- S16: Which disclaimer/warning on legal issues?
- S17: Are there any signs of sophisticated web page (flash animations, search options, web forums etc.)
- S18: What are the shipment options?
- S19: Does a web page contain any other links to sites with more / further / additional information, e.g. erowid.com, bluelightblulight.ru, other drug forums or other web shops selling NSD?
- S20: Is it retail or wholesale online shop?
- S21: When was the last update of the web shop?
- S22: What is the type of the web shop - RC shop/Commercial/Herbal/Other?
- S23: Was a web shop monitored in previous EMCDDA monitoring? If yes from which date?

PRODUCT

- P1: URL of web shop
- P2: Product name
- P3: Product chemical name
- P4: Product sold as? (legal high, herbal high, research chemical, bath salt, etc.)
- P5: Minimum quantity for sale
- P6: Maximum quantity for sale
- P7: Price for minimum quantity
- P8: Price for maximum quantity
- P9: Quantity measure (grams, liters, etc.)
- P10: Price currency (CZK, EUR, etc.)

- S11: Is there any information about product ingredients
- S12: Is there any information about any tests on the compound
- S13: Are effects of the product described?
- S14: Categorization of the product (e.g. stimulant, relaxant etc.)
- S15: Information of dosage of the product
- S16: Any promotion of extra/super experience relating to product
- S17: Product info - notification "new", "special offer", etc.
- S18: Product info - price trick - special offer, new price etc.
- S19: Product info - labeling - colorful/attractive labels
- S20: Product info - gift with the product, comparison with competition, etc.
- S21: Product info - in which forms a product is being sold?

2.3. Analysis of recent search technologies automation and optimization tools for primary web site search and data mining (products, prices, quantities etc.)

The technology used for automatic retrieval of data from the internet is called “scraping”.

Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox.

Web scraping is closely related to web indexing, which indexes information on the web using a bot or web crawler and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software. Uses of web scraping include online price comparison, weather data monitoring, website change detection, research, web mashup and web data integration.

SOURCE: Wikipedia

2.4. Selection/Creation of software tool for automation/optimization

New software (scraping tool) will be developed by CUNI for automatic search of web shops and detailed monitor of selected web shops.

Partners has to provide key words, by which an online shop is clearly identified in their local language (e.g. web shop, basket, delivery etc.) and according to this, the scraping tool for each partner will be set up.

Partners has to agree which shops to be monitored in more details for their country (this should be based on previous EMCDDA monitoring results, web shops which remain longer on the market and are most popular).

Manual for new software tool will be provided by CUNI to all partners. Language of the manual will be English.

The scraping tool will be provided by CUNI, same as the outcome analysis, however, the partners will be asked to contribute by coding particular national online shops into the scraping tool (after all the shops in the national country had been scraped and subjected to selection of the 6 key shops).

2.5. Testing of the new software tools

Testing will be done primarily by CUNI with co-operation of all partners with regard to language specifics issues and correction of software testing results.

2.6. Piloting, collection, processing and analysis of data

Piloting, collection, processing and analysis of data will be done primarily by CUNI with co-operation of all partners.

2.7. Report on most marketed substances online by all partners and overall, feeds into final Top 10 and technical folders on compounds

Report on most marketed substances online, feed into TOP 10 and technical folders will be done by all partners by equal distribution of required tasks.

2.8. Drafting of and communication on final report

Drafting of and communication on final report will be done primarily by CUNI with co-operation of all partners.

2.9. Preparation of a conclusion session on online shops at the third project steering committee/working group meeting

Preparation of a conclusion session will be done by CUNI during the last steering committee/working group meeting.

2.10. Preparation of article

Preparation of an article will be done primarily by CUNI with co-operation of all partners.

Year 1 report on the monitoring of online shops

Department of Addictology, First Faculty of Medicine,

Charles University in Prague and General Teaching Hospital in Prague

Co-funded by the European Union

Daniela Kmetonyová

Martin Pažitný

1. Introduction

During the first year of iTrend project implementation, we were able to gather following data describing the situation on the online market with new psychoactive substances thanks to newly developed methodology and software tools, which enabled automatic data gathering.

Data for 5 partner countries are included - Czech republic, Poland, France, UK, and Netherlands.

Main objective was to monitor online shops selling NPS (normal web, not „darknet“) which are relevant for each partner's country, what means that they have interface in local language or are relevant for the partner country by other means (data from fora analysis, survey, etc.).

2. Number of active web shops selling NPS

The table below shows number of monitored active web shops for 09/2014.

Number of active webs selling NPS

Country	Monitored domains
CR	25
FR	92
PL	97
UK	308
NL	23

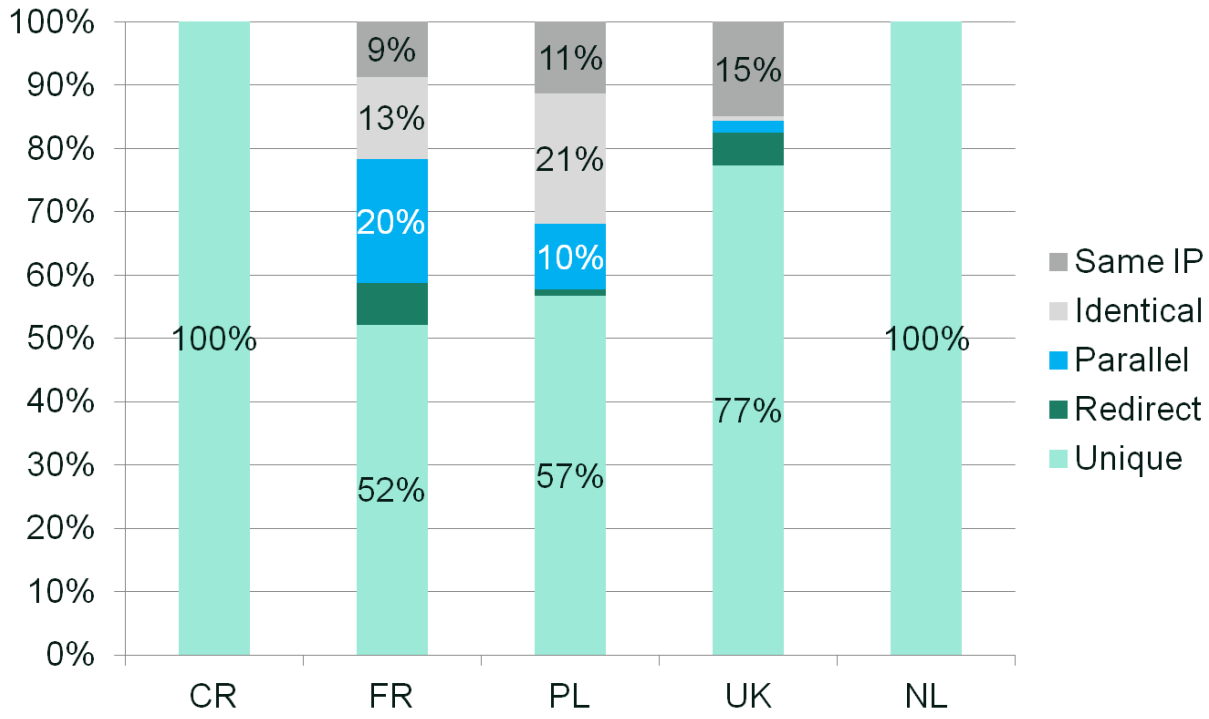
3. Active web shop categorization - Removing duplicities

Following graph shows % of all unique web shops in comparison to the duplicities which occurs on the market, thus reflecting more exactly true market size.

- **Unique shops** – web shops having unique design and IP address
- **Redirect** – web shop domains which redirect to another existing unique shop
- **Parallel** – web page which has the same design as existing unique shop (in category „Unique shops „we store the more popular from web shop)
  (Example: www.name1.cz, www.name1.eu, www.name1.com)

- **Identical** – web shop, which has same design and IP address than one already in „Unique shops „category.
- **Same IP** – web shops with same IP address but different design than web shop already in „Unique shops „category

Active web shop categorization – Removing duplicities, data for 09/2014



Limitation

- Only selected web shops were able to include for further monitoring by the 2nd software Shop Scraper due to the limited capacities

Monitoring is based on quantitative and qualitative indicators.

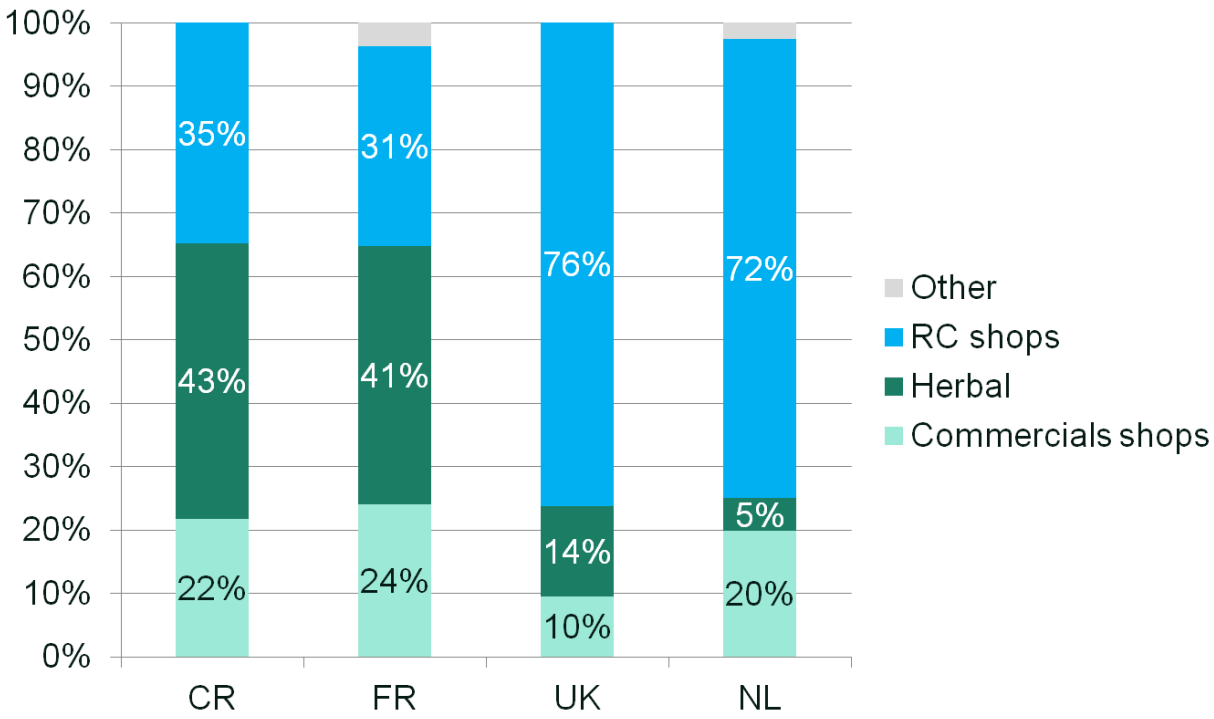
4. Active web shop categorization - Targeted population

Within our new methodology we introduced the following new categories of web shops selling NPS:

- **Commercial shops**
  - offering mostly branded products, sold under commercial names as "Funky", "Cocolino"
  - chemical compound of the substances are mostly unknown to the customer
  - it is often a “smartshop” or “headshop”
  - targeting less experienced users/wider population
- **2. “RC shop”, “RC”**
  - stands for “research chemicals
  - sells NPS mostly under their chemical names
  - targeting more experienced user
- **3. Herbal shop** – mostly selling herbal products such as kratom or salvia etc.
- **4. Other –** e.g. drugstores, web shop selling fitness food supplements etc.

Data for Poland wasn’t available at the moment of report creation.

Active web shop categorization – Removing duplicities, data for 09/2014

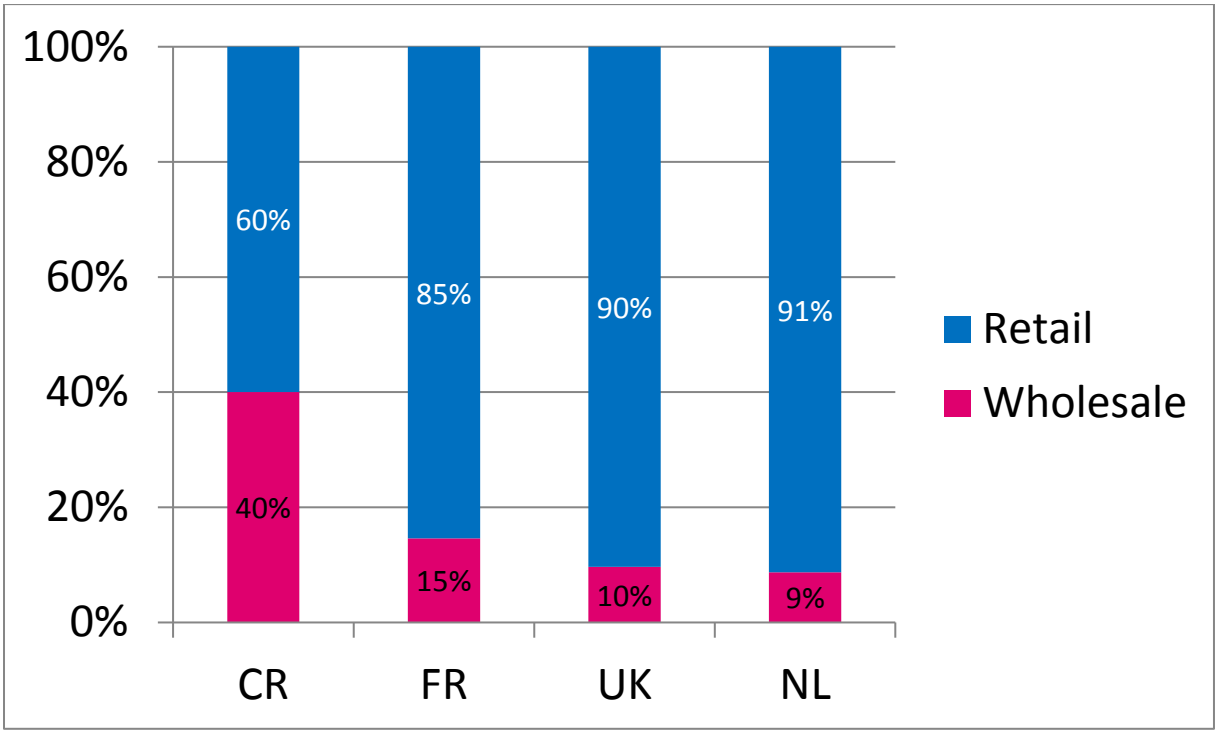


5. Active web shop categorization - Wholesalers vs retailers

Following graph displays the % of wholesalers and retails on the market.

Data for Poland wasn't available at the moment of report creation.

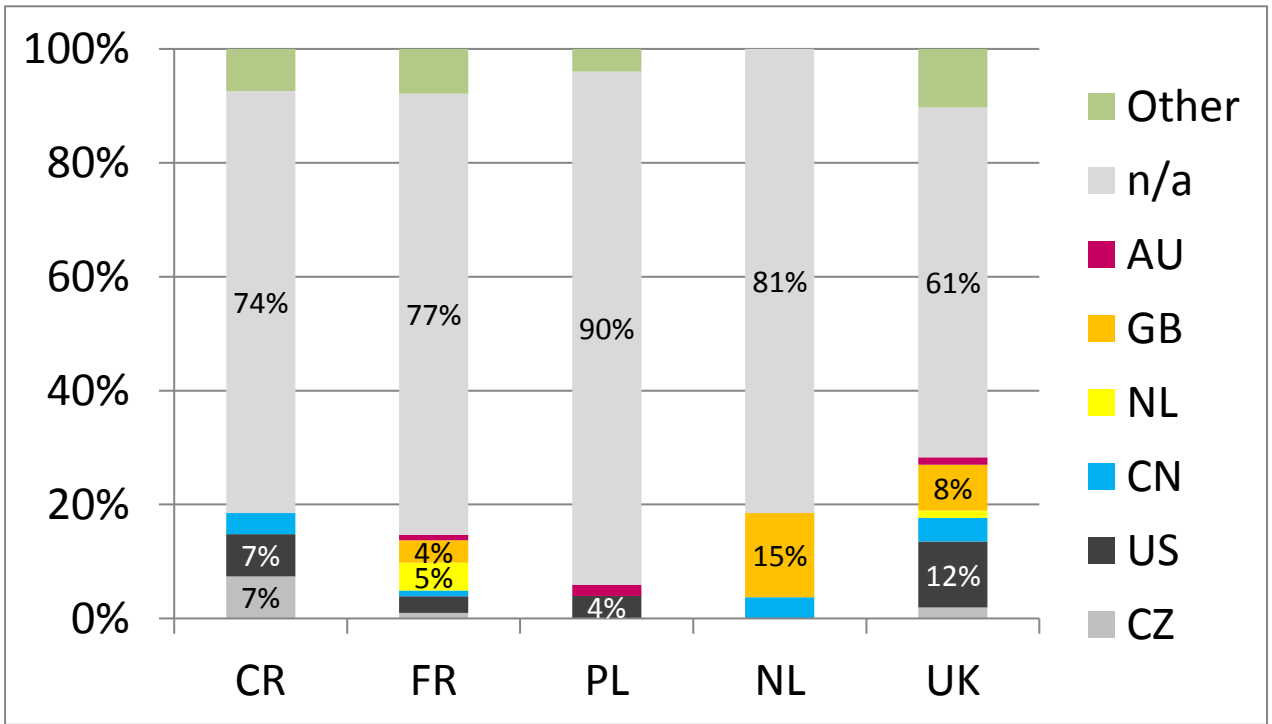
Active web shop categorization – Wholesalers vs retailers, data for 09/2014



6. Active web shop categorization - Location – Domain Owner Address

Following graph displays location of domain owner, thanks to which we are able to assess the potential location of the web shop.

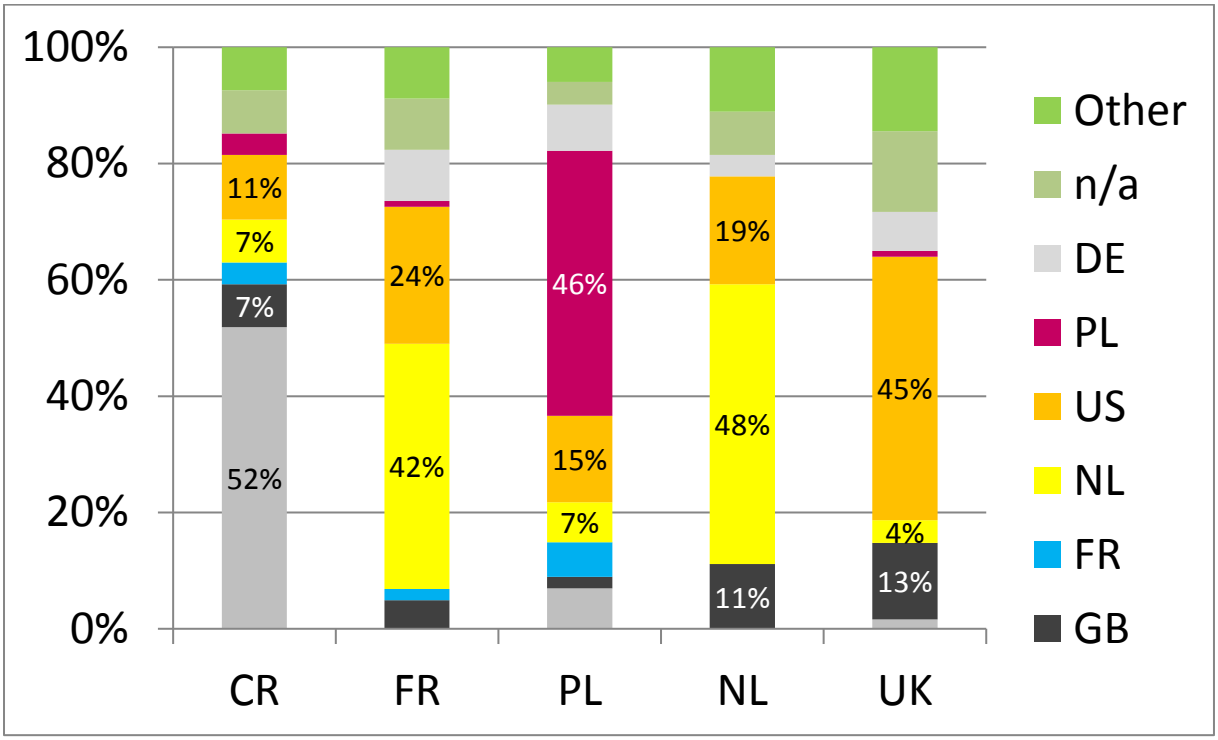
Active web shop categorization - Location – Domain Owner Address, data for 09/2014



7. Active web shop categorization - Location – IP Address Country Code

Following graph displays location of IP address, thanks to which we are able to assess the potential location of the web shop, mostly where the data from the web shops are stored and hosted.

Active web shop categorization – Location – IP Address Country Code, data 09/2014



.

Year 2 report on the monitoring of online shops

Department of Addictology, First Faculty of Medicine,

Charles University in Prague and General Teaching Hospital in Prague

Co-funded by the European Union

Daniela Kmetonyová

Martin Pažitný

1. Introduction

During the whole iTrend project implementation, we were able to gather following data describing the situation on the online market with new psychoactive substances thanks to newly developed methodology and software tools, which enabled automatic data gathering.

Data for 5 partner countries are included - Czech republic, Poland, France, UK, and Netherlands.

Main objective was to monitor online shops selling new psychoactive substances[2] on Surface web (not Deep Web), which are relevant for each partner's country, what means that they have interface in local language or are relevant for the partner country by other means (data from fora analysis, survey, etc.).

2. Number of monitored web shops selling NPS

The table below shows number of monitored web shops offering NPS.

	05/2014	08/2015
NL	27	38
UK	309	392
PL	103	139
FR	112	141
CR	33	41
TOTAL	**584**	**751**

3. Active web shop categorization - Removing duplicities

Following graph shows % of all unique web shops in comparison to the duplicities which occurs on the market, thus reflecting more exactly true market size.

- **Unique shops** – web shops having unique design and IP address
- **Redirect** – web shop domains which redirect to another existing unique shop

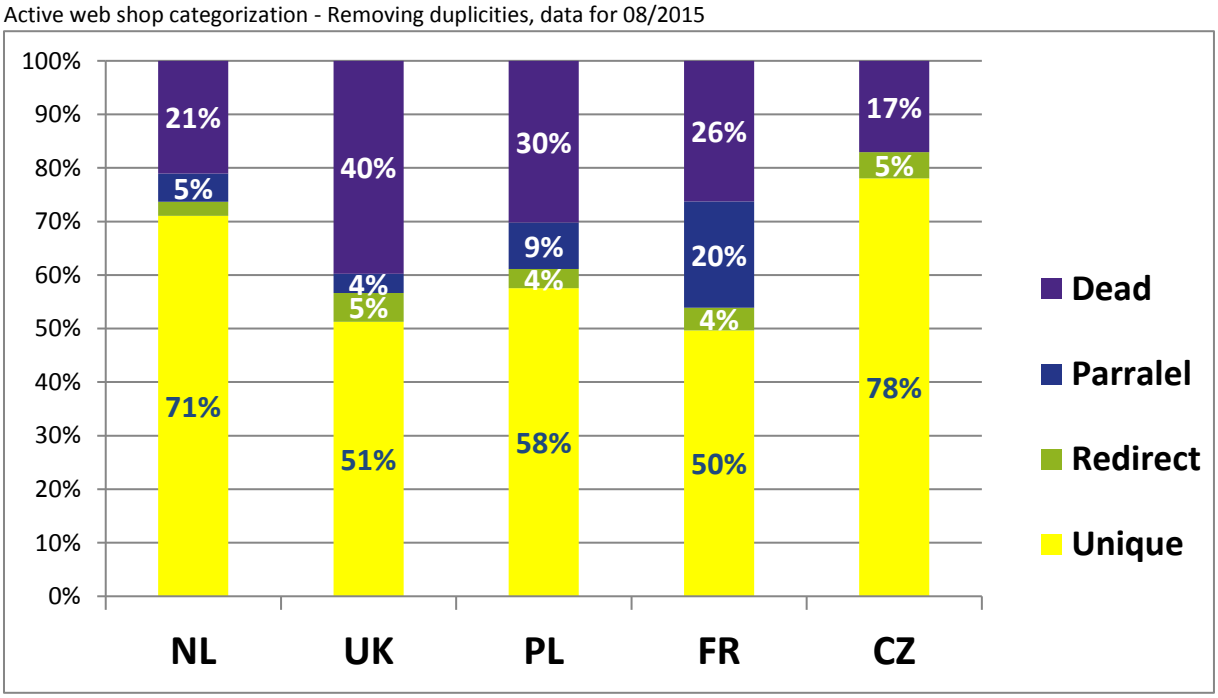
[2] New psychoactive substances - (EMCDDA definition) – 'a new narcotic or psychotropic drug, in pure form or in preparation, that is not controlled by the United Nations drug conventions, but which may pose a public health threat comparable to that posed by substances listed in these conventions'

- **Parallel** – web shops with the same graphic design but a different URL than an existing shop in „Unique shops“.
- **Dead** – Web shops, which are currently inactive

Number of Unique web shops

	Number of monitored web shops		Number of UNIQUE web shops		% of UNIQUE web shops	
	05/2014	08/2015	05/2014	08/2015	05/2014	08/2015
NL	27	38	19	27	**70%**	**71%**
UK	309	392	207	201	**67%**	**51%**
PL	103	139	72	80	**70%**	**58%**
FR	112	141	64	70	**57%**	**50%**
CR	33	41	30	32	**91%**	**78%**
Total	584	751	392	410	**67%**	**55%**

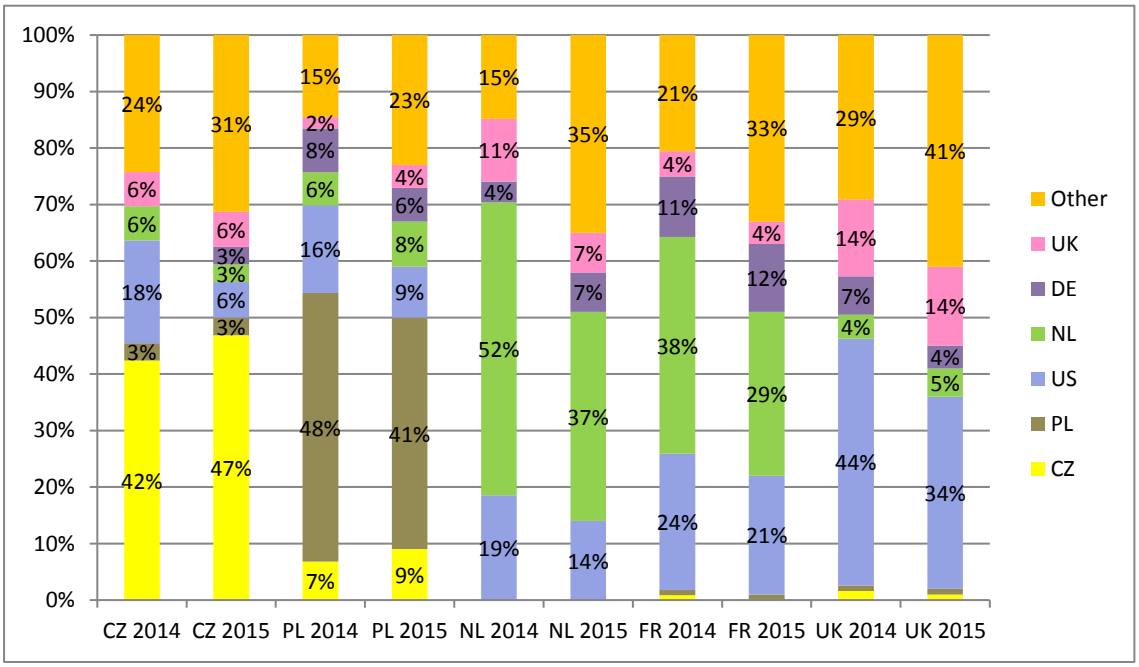
Active web shop categorization - Removing duplicities, data for 08/2015



4. Active web shop categorization - Location – IP Address Country Code

Following graph displays location of IP address, thanks to which we are able to assess the potential location of the web shop, mostly where the data from the web shops are stored and hosted.

Active web shop categorization – Location – IP Address Country Code, data for 05/2014 and 08/2015



5. Monitoring of selected web shops and their products

As part of the new methodology and also thanks to our second new software “Product scraper” we were able to gather the following information from selected web shops.

List of selected web shops for further monitoring and for first intention as selected online shops for purchase

UK	FR	NL	PL	CZ
chemicalwire.com (RC)	chemicalwire.com (RC)	drsmart.nl (RC)	kolekcjoner.nl (CO)	cs.buzz-wholesale.co (RC)
sciencesuppliesdirect.com (RC)	www.magic-mushrooms-shop.com (CO)	rschemicals.nl (RC)	legalchem.pl (RC)	mefedronprodej.webnode.cz (RC)
www.brc-finechemicals.com (RC)	www.officialbenzofury.com (RC)	www.rcnederland.eu (RC)	poland.am-hi-co.com (CO)	www.euphoria-smartshop.eu (CO)
www.coffeesh0p.com (CO)	www.plantfeedshop.com (RC)	www.sensearomatics.eu (RC)	rc-lab.cz (RC)	www.plantfeedshop.com (RC)
www.discofood.com (RC)	www.shayanashop.com (CO)	www.wapwinkel.com (CO)	researchchemicals.net.pl (RC)	www.rc-chem.eu (RC)

Table below displays

a) Total number of individual items offered by all the shops for partner country (e.g. mephedrone in tablets and mephedrone as powder counted as 2 items)

b) Total number of unique products offered by all the shops for partner country (e.g. mephedrone in tablets and mephedrone as powder counted as 1 unique product)

c) 1st most often offered substance (from number of unique products) from all the shops for partner country

d) 2nd most often offered substance (from number of unique products) from all the shops for partner country

e) 3rd most often offered substance (from number of unique products) from all the shops for partner country

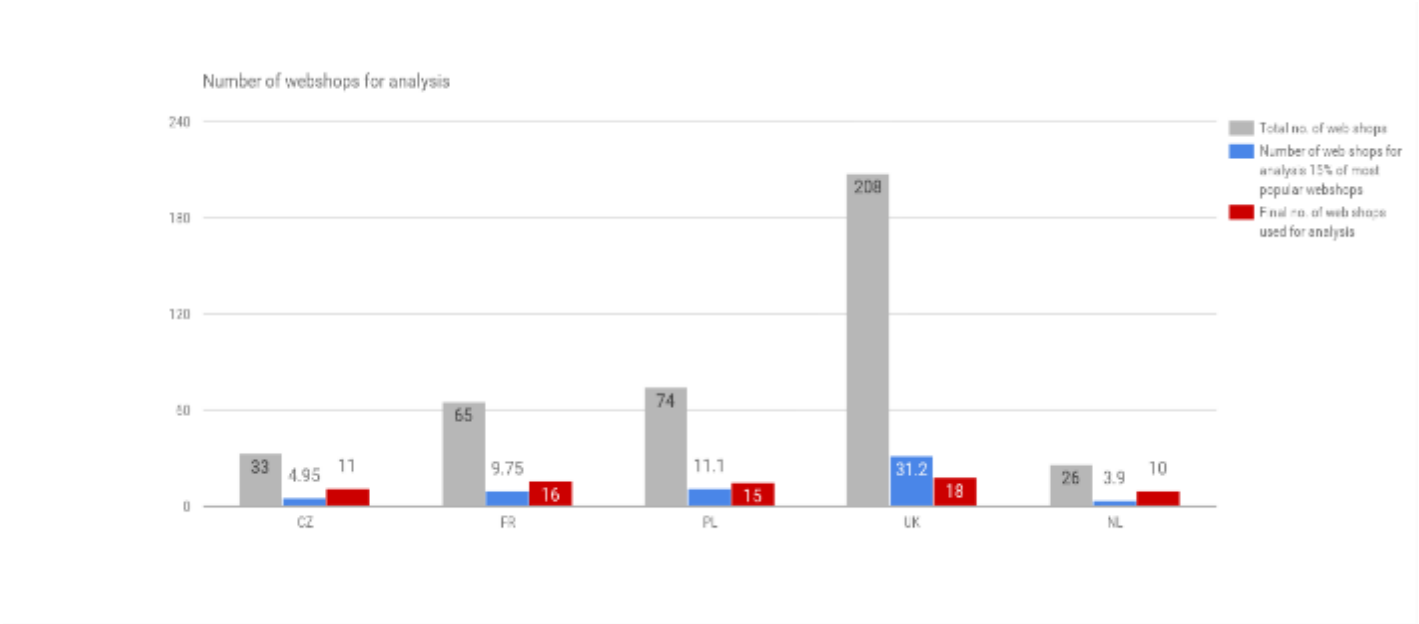
Data on selected web shops –number of offered items, unique product and most offered substances, data 08/2015

	a) Number of individual offered items	b) Number of unique products	c) Top 1st substance	d) Top 2nd substance	e) Top 3rd substance
UK	368	137	Ethylphenidate	BK-2C	MPA
FR	345	273	Salvia	**Kratom**	BK-2C-B
NL	844	304	**4-MEC**	JWH-122	4-FA
PL	317	256	Pentedrone	**3-MMC**	ETH-CAT
CZ	103	79	**Kratom**	**3-MMC**	**4-MEC**

6. Analysis of marketing strategies

The data for analysis of marketing strategies were collected for in several different time periods for different partner countries:

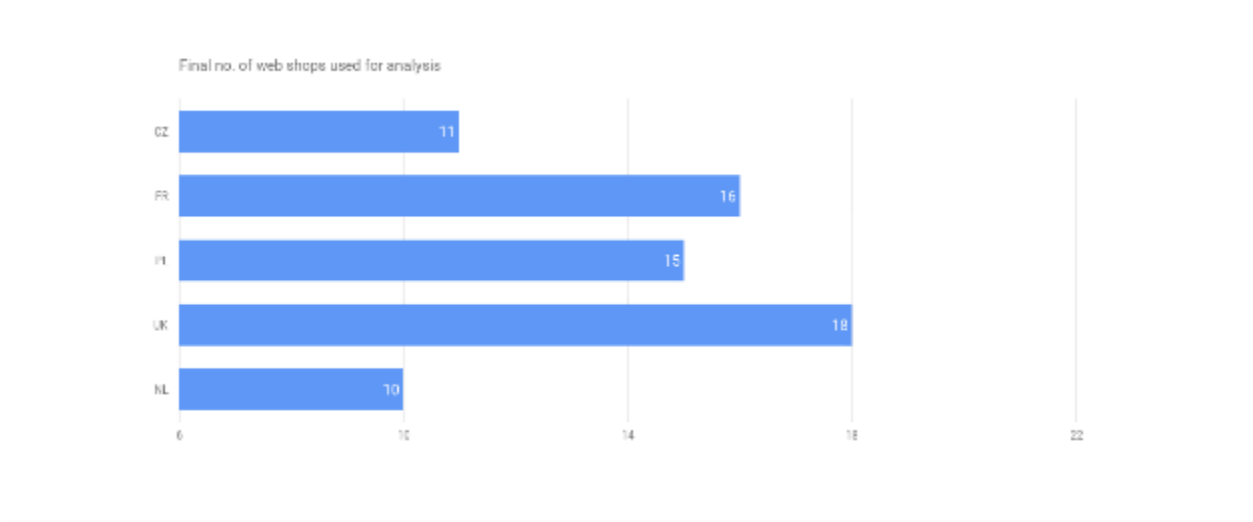
- UK in period between 28.3 2015 - 3.4. 2015
- CZ, PL and NL between 28.5.2015 - 1.6.2015
- FR between 28.5.2015 - 24.6. 2015.

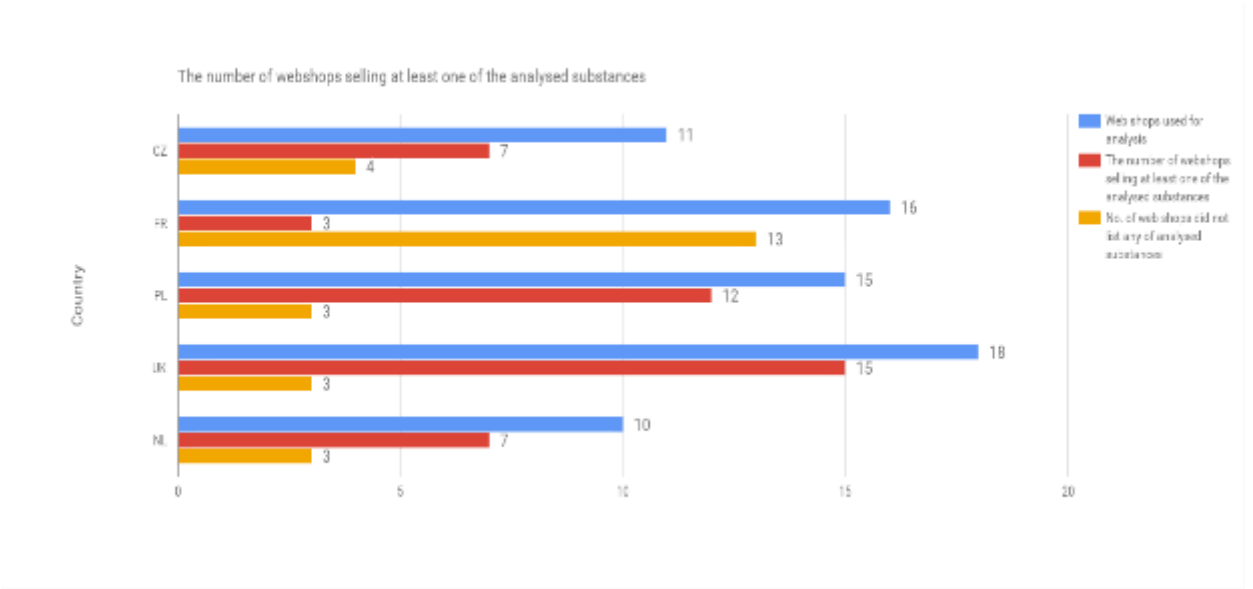


The first bar displays the total number of web shops selling NPS for partner country to the date 16.3.2015.

The second bar shows 15% of the most popular web shops relevant for partner country, which were selected for further analysis of marketing strategies.

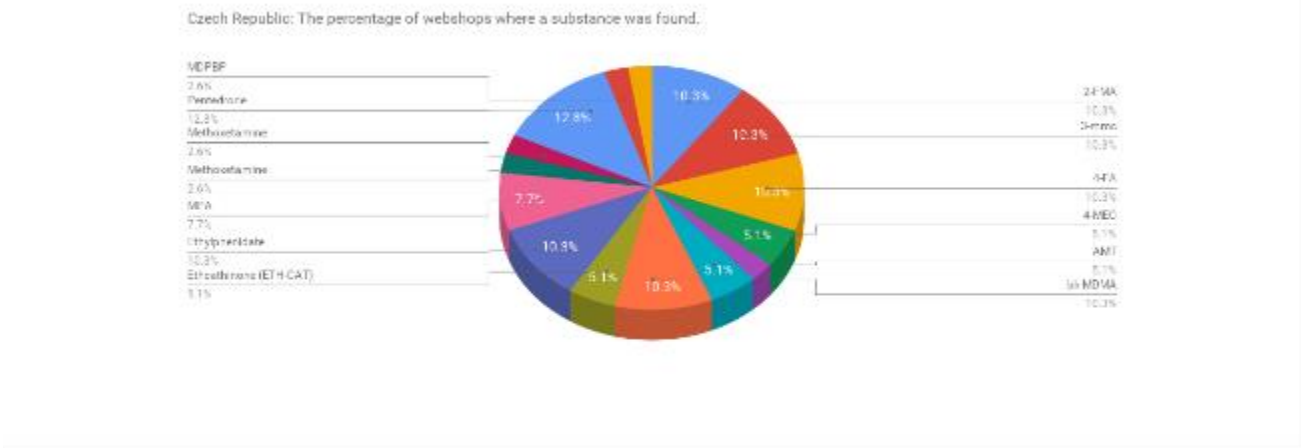
The third bar shows the final number of web shops which were used for marketing analysis. This number is different from the second bar, because several web shops were removed from this top 15%, e.g. the web shops operating on platform otwarte24.pl and shops which provide just bulk/wholesale order. In addition to that, another 6 web shops, previously chosen by the partner's countries for the purchase of substances (from 21.3.2014), were added to the final list of web shops for further marketing analysis (See paragraph 5).



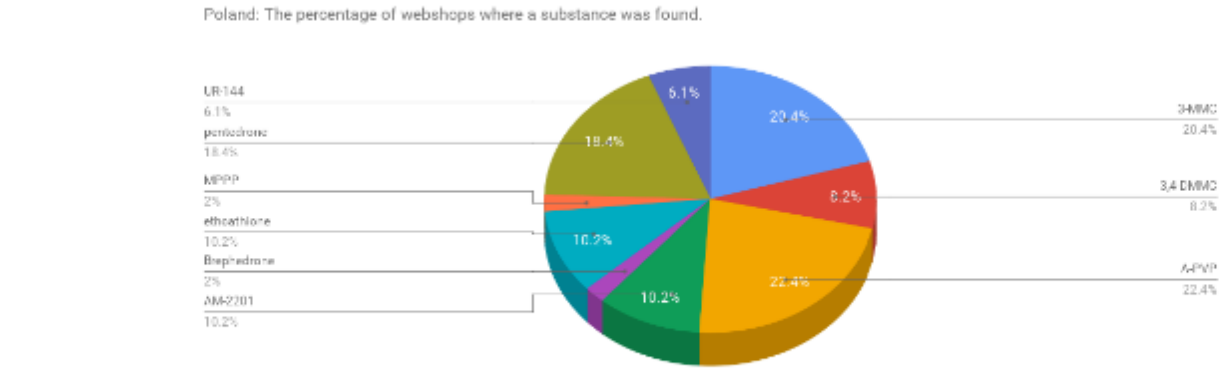


The marketing strategies were then assessed in relation with several substances for which National Technical Folders were prepared. For each country different number of substances was used (for details see the graph above).

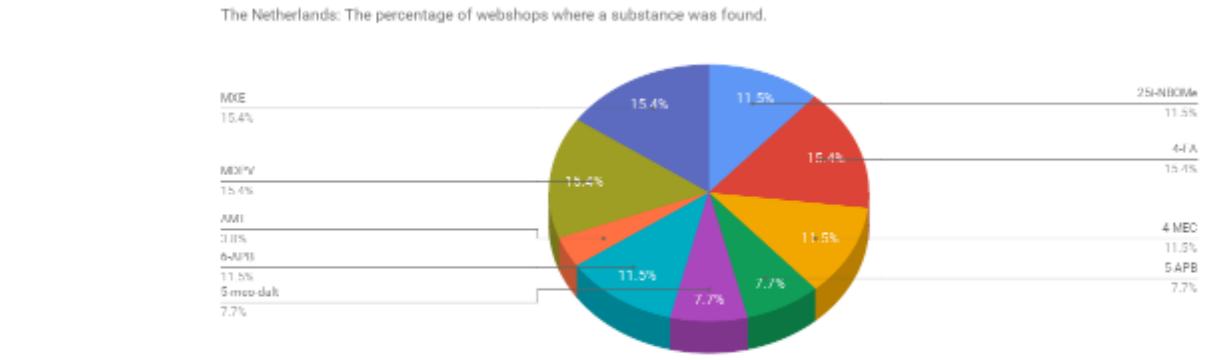
The graph below shows the offer of each analyses substance from selected web shops for each partner country. Included are also web shops, which were at that moment having status “Out of stock” for this substance.



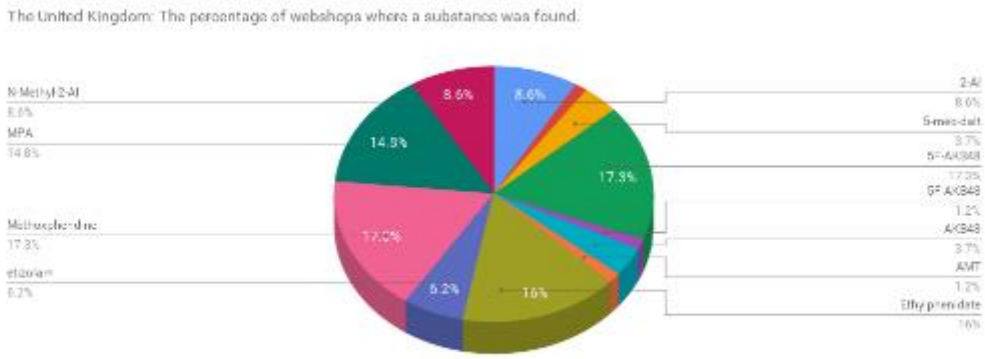
Another two analysed substances MDPB, PMPPP were not offered by any of the web shops.



Another three analysed substances MDPB, Mephedrone and Brephedrone were not offered by any the web shops.



Any of the web shops for the Netherlands offered 5-(2-aminopropyl)indole.



For France, only one substance out of all analysed substances was found in the shops – Ethylphenidate.

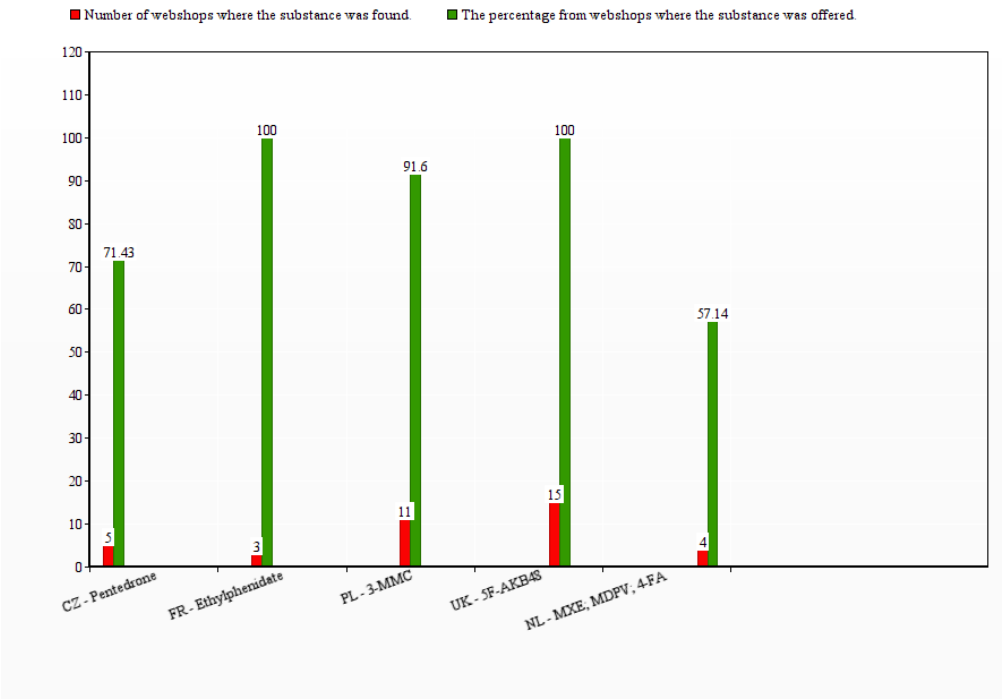
Two points can explain this particular observation for France:

- The selection methodology of websites for their marketing analysis was intended to retain the ones potentially frequented by the largest number of consumers. The addition of trade names in the Top List allowed to highlight the unsuspected importance for France of the "commercial websites". These sites sell predominantly either of plants, or a wide variety of commercial products as well as the paraphernalia. Indeed, retail sites that have been selected as the most popular with the French were not primarily sales sites specializing in the NPS (so-called RC shops) while the selection of the Top List NSP had been dictated by information that promoted the superiority of chemical names - so specialized sites - on names and commercial sites. But the workstream 2 objective was to select webshops on the criteria of the largest access on the broad type of consumers.

- The United Kingdom has excluded of its Top list the NPS which have been classified during the project, conversely, the France has not excluded from its Top List NPS which have undergone classification, either in his country or in nearby countries. The fact that during the project France could purchase NPS although considered in diffusion, could be a demonstration of the will of the sites of online sales to show a respect for the laws in force during the time between the establishment of the Top List, the online purchases and the monitoring of the commercial offer of the selected websites. Also a last hypothesis is that legislative decision-making from the USA and the United Kingdom may be those that have the greatest impact the availability of products on the Internet from surface.

For the purposes of the activity 2 and 3 should have been to create two lists, one established under activity 5 representative considered NPS streaming, the other listing the NPS the most streaming but not subject to a classification in France, the United States and the United Kingdom. This last list could better allow the exploration of their availability on the Internet.

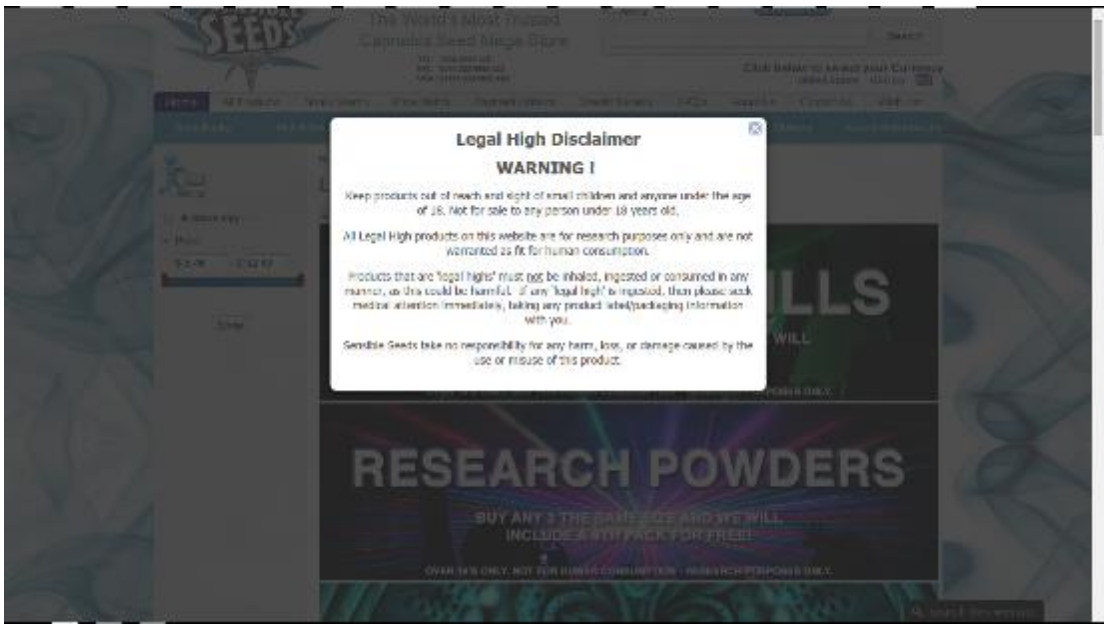
Magali Martinez, Agnes Cadet-Taïrou (OFDT)

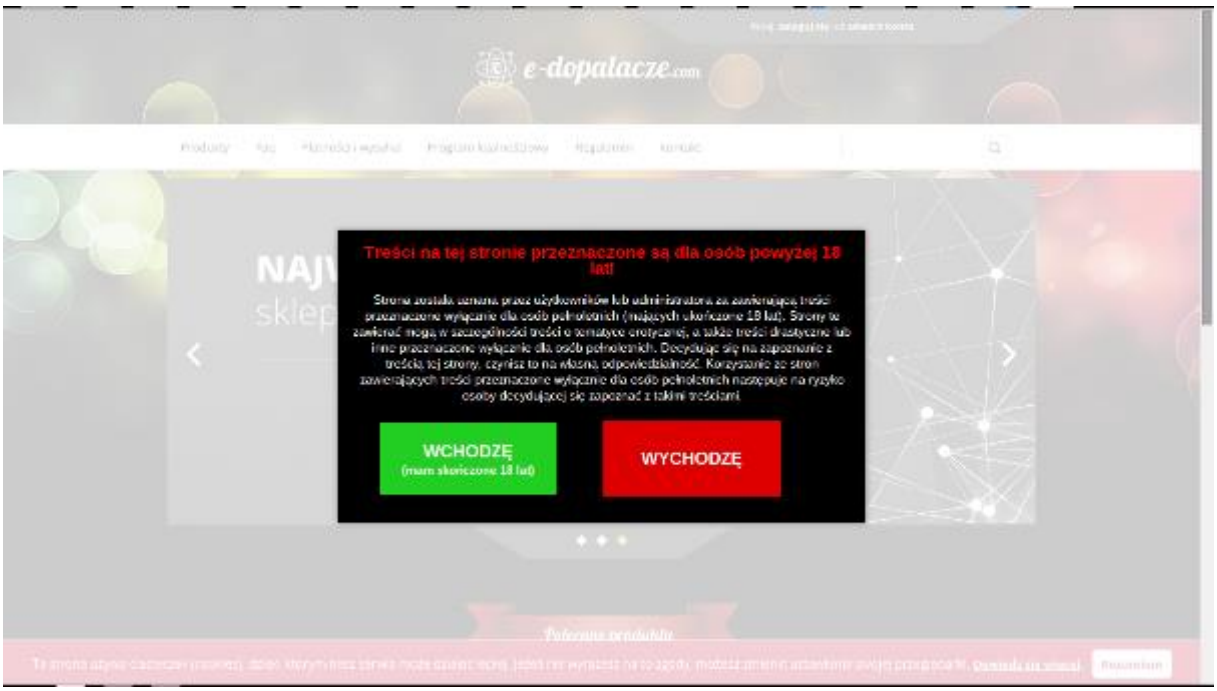


7. Analysis of legal notice – Age restriction

As a part of monitoring activity, an analysis on age restrictions present on selected web shops was performed. 48 most popular web shops from the partner's countries were selected. The analysis found that 38 web shops used some form of age restriction on their pages and identified 3 different ways of informing about age restriction.

The most prevalent way of informing about age restriction was providing this information as a part of "Terms and conditions" or "FAQ". This way was used by 22 web shops out of 38 (58%). The second most prevalent way was providing this information on web shop main page (3 shops/8%) or via pop-up window (4 shops/11%). Examples of pop-up windows from analysed web shops are below:





The third strategy was informing about age restriction as a part of substances' information. This strategy was used by 6 web shops. Five web shops used small graphical icon located on picture of a substance, as you can see below and one web shop used text description.

Another 10 web shops didn't not provide any information on age restriction at all.

The table below shows summary information on age restriction analysis of selected web shops for partner countries.

Country	No of web shops	Information as a part of Terms and condition	Information on main page	Information provided via pop-up window	Product page – Graphical icon	Product page – Text description	No information at all
CR	11	7	0	0	2	0	2
FR	10	7	0	0	1	1	1
PL	8	2	2	3	0	0	1
UK	12	5	1	1	3	0	2
NL	7	2	1	0	0	0	4
SUM	**48**	**23**	**4**	**4**	**6**	**1**	**10**