Machine learning algorithms and the usage of domain names. How does a robot see the .pl domain?

Piotr Studziński-Raczyński
Senior DNS Specialist

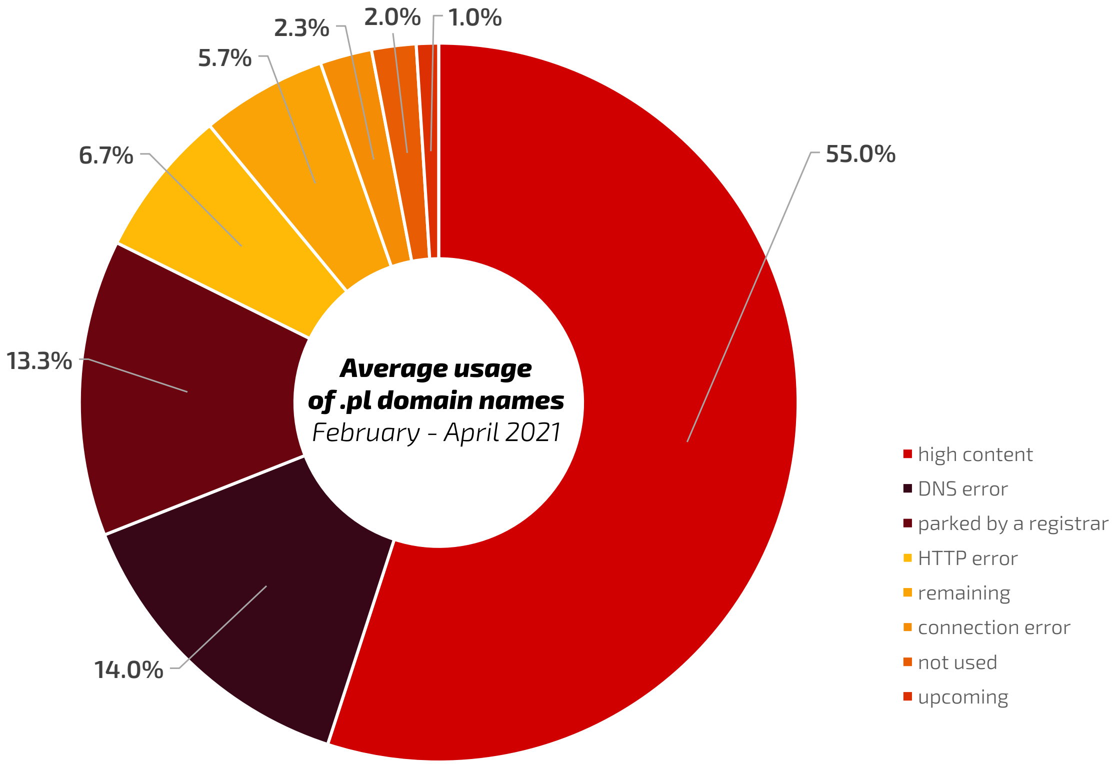
In order to examine and categorise the usage of Internet domain names, the Internet robot ‚crawler' *signs of life* was developed within the CENTR work as an indexing tool, allowing monthly data to be collected on the basis of a random sample (50k) of domain names, provided by the participating members of this initiative, including the .pl domain registry.

The method of usage of domain names is important information, allowing to better understand what happens with the registered domain names, e.g. to know the scale of occurrence of particular errors resulting in incorrect service delivery or to get an insight into what part of domain names the registrants have not found application for yet. The resulting picture can help build strategy, set goals, assess business risk and influence the outcome of decisions. The percentage of domain names „parked", with errors, redirected or presenting websites with relevant content can give an overview of the general interests and needs of registrants, help research current trends in the domain market as well as support the efforts of the registry, registrars and domain name users themselves to make the most efficient and adequate use of web resources.

The *signs of life* platform, based on a machine learning algorithm, scans and classifies the input in defined technical areas. This helps to determine how much of the top-level domains are used to present contents on the Internet. Thus, the classification is divided into categories in terms of the granularity. At the first level of detail, we distinguish whether the examined domain name refers to any contents. Domain names, not related to any contents, are those, the operation of which (website loading) is blocked due to some error, e.g. DNS error, connection error (server error, timeout) or http error with a specified code. Domain names referring to specific contents have been, in turn, divided into those presenting the high content and those containing information on their category (low content), e.g. „parked" by a registrar, blocked, not used, upcoming, abandoned. The last, fourth level of detail, specifies the category of domain names defined at the previous level, which in some cases is related to further and final subcategorization. An example would be the subdivision of unused domain names into those referring to a site index or to a blank page. Another example is the further subdivision of domain names with the category upcoming into those presenting the information „page under construction" and those presenting the initial page of the website builder.

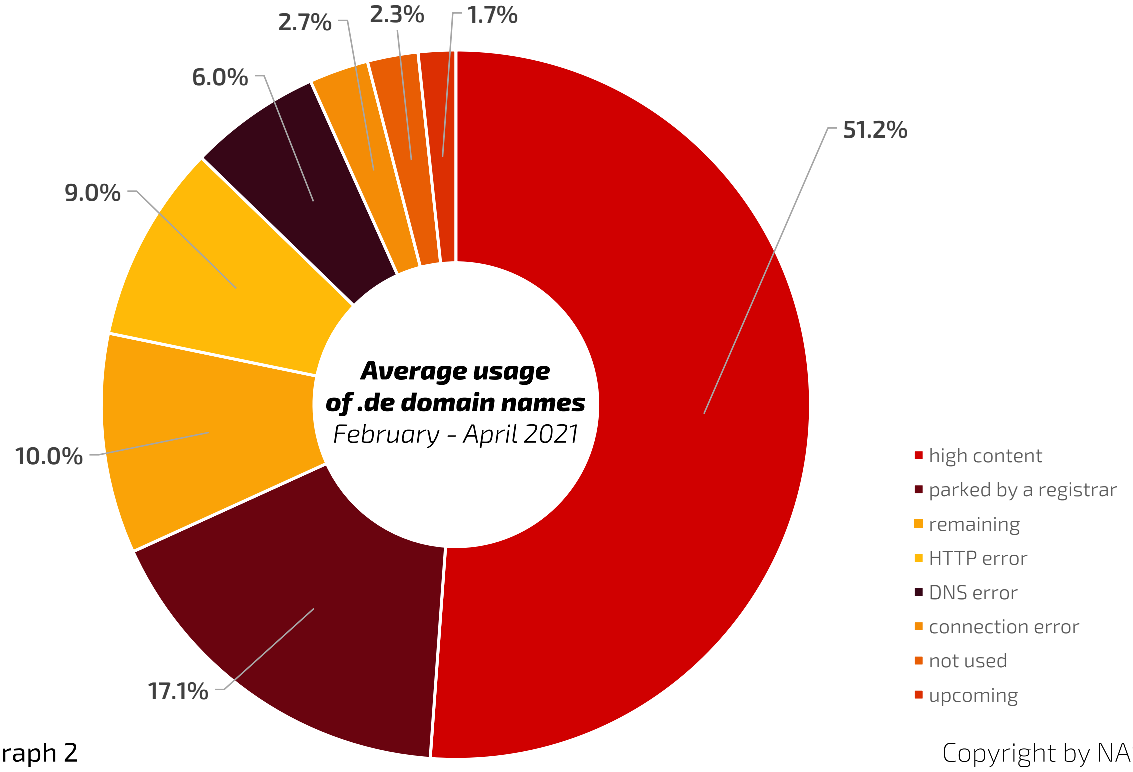
Besides examining domain names for their contents, the crawler verifies the extent to which they are redirected to external addresses, as well as how many of the domain names analysed are linked to an email service - if a domain name fails to return a website, this does not necessarily mean it is not in use. It may also be associated with a mail server. To get this information each domain name is queried for MX (Mail Exchange) records using the dnspython library. For example, in May 2021, the MX record index for the .pl domain, similarly to the .fr domain, was 88% and it was the second highest result after .de and .si domains, where the Mail Exchange service was found in 89% of the examined domain names. To compare, the average value of this indicator for other ccTLDs, subject to the analysis, amounted to 76%.

On the graph no. 1 below, average data on the usage of .pl domain names, for the period from February to April 2021, has been presented. Corresponding data for the same period have been presented on the graph no. 2 for the .de domain. On the basis of the data obtained it may be observed that the percentage of domain names referring to the contents, where the high content is presented, in the case of both analysed domains is similar and exceeds 50%, with a few percentage points advantage for the .pl domain registry. These results may prove a good effectiveness of usage of the domain names registered by the registrants. Another criterion that may draw an observer's attention is the occurrence of DNS error, i.e. incorrectly configured domain name delegation. In this respect, it can be noticed that names, registered at our western neighbours, more often have the correct delegation indicated.



Graph 1

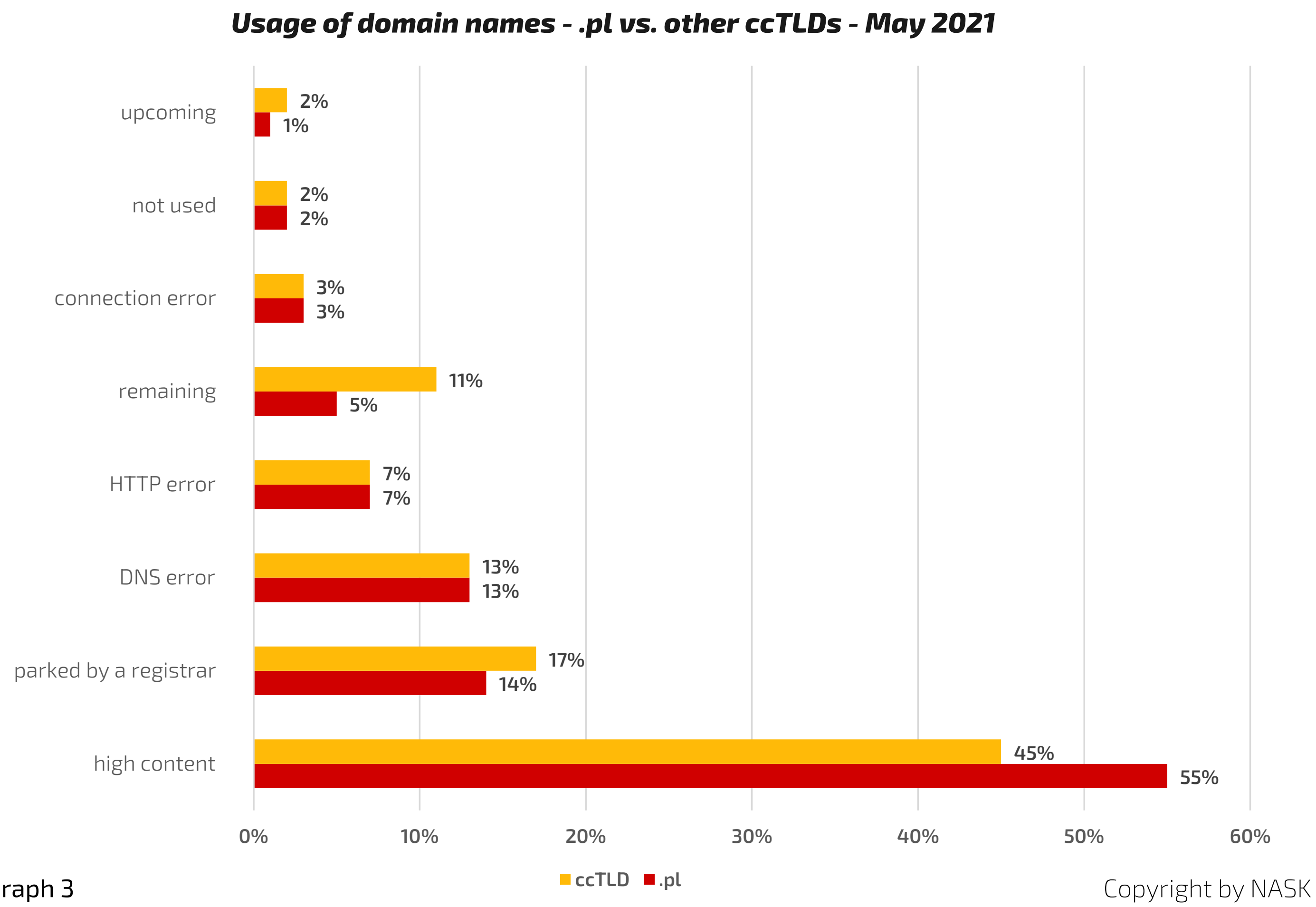
Copyright by NASK



Graph 2

Copyright by NASK

On the other hand, looking at a more global level, i.e. comparing the results, obtained in May 2021 for the .pl domain names, with the average values for other country code top level domains, taking part in the research (graph no. 3), one may observe that the percentage of effective use of .pl domain names, amounting to 55%, considerably exceeds the average for the ccTLD registries, covered by the research, where this value remains at the level of 45%. The .si domain is an example of an outstanding result in this category, where the value of the index reached 60%. At the same time, it is also worthy of note that the Slovenian country code domain records a low rate of domain names „parked" by registrars (3%) as compared to the average for other national domains equalling 17%. In May 2021, for the .pl domain that figure was kept at the level of 14%. In case of the remaining categories, covered by the research, the values noted by the .pl domain were equal to or close to the average obtained by other ccTLDs.



Graph 3

Copyright by NASK

There is a view among some representatives of European registries that the possibilities, offered by the crawler, constitute a first step towards a more detailed characterisation of the content to which Internet domain names refer. It is also a common view as to the merits of machine learning-based analysis and the multilateral benefits of the information obtained in this way. Such analyses are already being conducted or are in the pipeline. Pointing out the pros of this kind of zone mapping, e.g. those related to the use of the obtained data to assess the strength and quality of the zone, even from purely technical aspects, the representatives of the ccTLD registries at CENTR emphasise, at the same time, the importance of a rational evaluation of the data, thanks to which it is possible to draw conclusions allowing for an even more effective and safer use of the possibilities the domain names offer to the Internet users.