Using webcrawler techniques for improved market surveillance – new possibilities for compliance and energy policy

Peter Bennich
The Swedish Energy Agency
Box 380 69
100 64 Stockholm
Sweden
peter.bennich@energimyndigheten.se

Signe Friis Christensen
The Danish Energy Agency
Amaliegade 44
1256 København K
Denmark
sfc@ens.dk

Kasper Mogensen
Big2Great ApS
Lupinvej 19 st.
2720 Vanløse
Denmark
ksm@big2great.dk

Linn Stengård
The Swedish Energy Agency
Box 380 69
100 64 Stockholm
Sweden
linn.stengard@energimyndigheten.se

Teemu Hartikainen
Finnish Safety and Chemicals Agency (Tukes)
33100 Tampere
Finland
teemu.hartikainen@tukes.fi

Troels Fjordbak Larsen
Big2Great ApS
Lupinvej 19 st.
2720 Vanløse
Denmark
tfl@big2great.dk

Keywords

market surveillance, monitoring, ecodesign, energy labelling, webcrawler data

Abstract

In recent years, web crawler techniques have been developed at a rapid pace. With intelligent software, it is possible to scrape large volumes of data from publicly available data sources at the Internet, basically in real time. In the realm of product policies, such as ecodesign and energy labelling, this offers *alternative means* to *track* products available at the marketplace, when compared to more traditional data collections offered by companies like GfK. Thus, typical information on products and model names, claimed performance on functionality and energy use, and finally purchase price, is thus affordable and relatively easy to collect in real time. With advanced modelling based on ranking (popularity indexes at sites like e.g. Pricerunner) it is even possible to estimate the sales volumes.

The new possibilities thus offered are currently being explored for policy evaluation and improved policy design, as well as for developing new and improved consumer tools when choosing and buying energy using products.

This paper focus on another use, namely on how webcrawler techniques can increase the effectiveness of *market surveillance*. In the "NordCrawl" project, under the Nordic Council of Ministers, market surveillance authorities (MSAs) from the Nordic countries work together to develop a web-based application which each MSA can use to *monitor* their respective national market.

The advantages include the following aspects:

- Reduced cost for data collection
- Higher coverage and a better representativeness of the market (market picture)
- Data collection in real time
- Improved sampling strategies when choosing products to check, be it random or targeted sampling
- Faster response times for targeting non-compliant products
- Overall improved effectiveness and reduced cost for market surveillance

This is a new way of working for MSAs and the paper discusses initial lessons learnt. Apart from the advantages above, it also includes issues like retailer acceptance and legal aspects.

Introduction

The ecodesign and energy labelling regulations are two of the pillars of the energy efficiency measures in EU, providing almost half of the energy efficiency goal of 20 % to 2020 (Kemna, 2015). As is well known, however, the confidence in and compliance of the regulations stands and falls with an active, extensive and effective market enforcement, a burden put on the individual member states (MS) in the EU. Since this in general is considered a resource intensive activity, the degree to which individual MS do enforce regulations vary quite substantially. Considering the increasing number of products that are regulated, the cost and time will only increase, which calls for improved and more effective methods.

A critical component of enforcement, be it product testing or documentation control, is a *knowledge of the market itself* when choosing products: What products are sold, and how many?

When were they put on the market and what performance do they have? And how does the market evolve over time?

The larger the knowledge of the market, the larger the degree of freedom when choosing a suitable *sampling strategy*:

- High degree of knowledge of the distribution of *products* that are sold, including claimed performance. It also includes information on how that distribution moves along the value chain manufacturers – wholesale - retailers. This allows for a choice between two strategies:
  - *Random* sampling (in a statistical sense): Gives a good picture of the market. The results can be scaled up and used as an estimate of the compliance rate and hence to what degree the market adapts to the regulations or not. It can also provide aggregated estimates of lost savings due to non-compliance.
  - *Targeted* risked-based sampling: Can be used when it seems like claimed performance don't meet eco-design minimum requirements and/or if energy labels are missing.

In either case, one can choose not only product or product group, but also combine that with a particular actor along the value chain – e.g. pinpoint a cluster of wholesales or specific type of retailers etc.

- Low degree of knowledge of the market: In this case, products can only be chosen *ad-hoc*. This is not effective since neither one of the benefits of random and targeted sampling can be achieved. However, this is probably the typical case for MSAs without enough resources to monitor the market.

Concluding that a detailed knowledge over the market is key for an effective market surveillance, the question is how that can be obtained. Typically, high quality market data from institutes like GfK[1] can be bought, but are expensive, especially if a high degree of details is required. Sometimes MSAs from different countries (or within countries) cooperate to share the burden of data gathering and enforcement, which can lower the cost, but might increase the administrative work load.

For all the reasons above, MSAs from the Nordic countries are currently working together in a project to investigate whether *web-scraped data* can be used instead, that allows each MSA to *monitor* their respective national market. The project is called "NordCrawl" and runs in parallel with the Nordic cooperation on Market Surveillance, "Nordsyn" (Nordsyn), under the Nordic Council of Ministers.

The paper starts with an introduction to the concept of web-crawling and what kind of data that can be collected. Then legal issues and a comparison with traditionally purchased data is discussed. Finally, some results are shown for a few products as an example of the type of information that is possible to get, and how that can be used for market surveillance purposes.

Basic elements of web crawling techniques

STRUCTURE OF THE NORDCRAWL SYSTEM

Figure 1 shows the basic architecture setup for the NordCrawl system. The system is divided into four different parts:

1. Information on products, stored at various publicly available web sites.
2. The web crawler engine and a temporary raw data storage.
3. A product data base containing processed data.
4. The analysis module (a program), which access and display the data in various ways.

HOW IT WORKS

In short, the web crawling process consist of two parts. The first part is to download the webpage and the second is to extract information from that page. The crawler is set to a start page (URL), from that page it will visit and download all linked pages until it reaches a specified depth (links away from the start page). When a page is downloaded, the information can be extracted using a source and appliance specific "recipe". The recipe specifies which information to extract as a specific product attribute. The recipe is either a specific address like "second row, first column in the table is the model name" or relative like "next word after 'model name:' is the model name".

In the third step, the raw data is processed and cleansed from redundant information; model names are aligned (see next section); finally, processed data are organised and stored in a database ("NordCrawl database").

In the fourth step, the data can be accessed from an application and where different aggregations and analyses can be done. The application is run by the MSAs and is in practice the tool they can use for monitoring the market.

The different parts in Figure 1 may be managed and owned by different actors. For instance, the actual crawler engine might be owned by a company (which can be procured under competition) whereas the raw data storage can be managed by the same company but owned by the MSA. The database with the final results and the associated application can both be managed and run by the MSA. It will probably be a combination of legal issues and convenience that finally determines the final management and ownership of the system.

WHAT KIND OF DATA CAN BE COLLECTED?

The amount of extracted data is indeed very large, as shown in Figure 2, which is why the processing in the next step is crucial. Product type, model name and number; different technical specifications and energy performance data have to be collected, analysed and grouped in a systematic way.

The model names constitute a particular hard challenge, since the same model may have alternate names in different countries, but as described below (Sec How does web crawled data compare with traditionally purchased data?) it is possible to get around.

As a guide to the way different models are characterised, the glossary in Figure 3 is used. With this as a basis, all the different technical and other performance data can be associated with individual models.

[1] Gesellschaft für Konsumforschung; see http://www.gfk.com/.

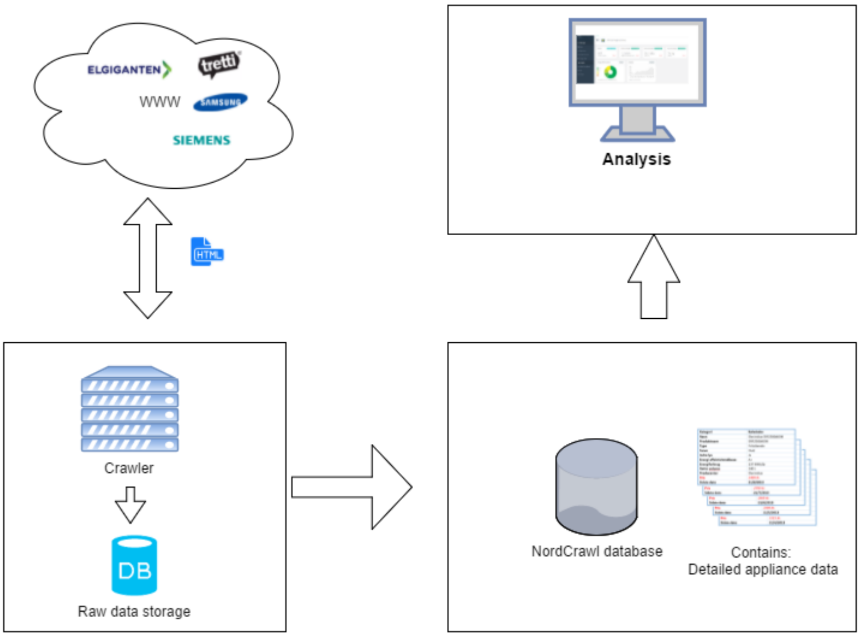


Figure 1. NordCrawl architecture.

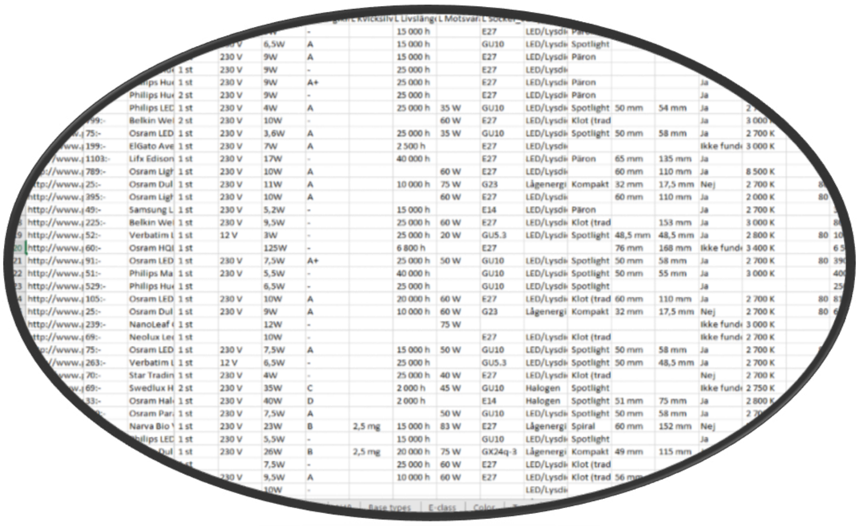
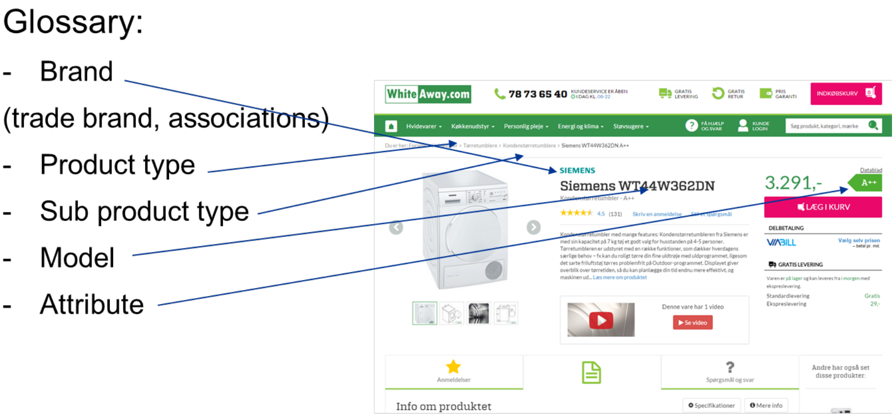


Figure 2. Data comes in large quantities and with a lot of details, here an example for lighting products.



Glossary:

- Brand

(trade brand, associations)

- Product type
- Sub product type
- Model
- Attribute

Figure 3. Glossary used for characterisation of products.

LEGAL ISSUES

All the data are collected at publically available websites and at first glance it should be of no concern. However, questions have been raised regarding a number of questions, such as:

- What legal possibilities and obstacles must be considered when collecting data ("crawling") from e-commerce websites that fall within the scope of the Nordcrawl tool?
- What legal possibilities and obstacles must be considered regarding storage of data collected by national regulatory authorities from e-commerce websites?
- What legal possibilities and obstacles must be considered regarding the national regulatory authorities' use of the collected data?
- Who shall be considered as the owner of the data when stored at the national regulatory authorities and what are the consequences?
- How does the principle of public access to official records affect the collection, storage and use of the data?
- What legal possibilities and obstacles must be considered regarding access to the information for other Nordic regulatory authorities?
- Can legally binding decisions for economic operators to remove products from marketplace be based on crawled data?

An initial legal assessment for the Swedish and Norwegian markets (SEA 2017) showed that it is basically deemed lawful to crawl data in the way described above and for this purpose, as long as no copyrights are violated and that the gathered information is made accessible to the public on request[2]. For the latter, it is important to remember that the NordCrawl only serves as a tool for market surveillance but that it isn't part of the formal procedure – that is an entirely separate process. However, there might be other reasons for protecting some of the processed data and analyses, in particular when it comes to results that might cause individual companies harm if made public. Thus, we anticipate a case-by-case assessment before a practice is developed.

ACCEPTANCE BY RETAILERS AND OTHER ACTORS BEING CRAWLED

Market surveillance is part of the legal framework within EU and in principle any economic operator, such as manufacturers, retailers and shops etc that put products on the market, have to provide MSAs information on request. Nevertheless, we believe it is important to have a dialogue with these actors if web crawling techniques are going to be used on a larger scale. Apart from informing them that they are under observation, it can also consider more practical things like when it is best to crawl the websites to disturb the data traffic the least; e.g. in the middle of the night.

HOW DOES WEB CRAWLED DATA COMPARE WITH TRADITIONALLY PURCHASED DATA?

If web scraped data truly are to replace purchased data (such as from GfK), it is essential to understand whether scraped data can provide at least the same information. For this reason, an analysis of and comparison between purchased and scraped data was made for three product groups, dishwashers, washing machines and TVs, over a time interval of a couple of years. We compared model names from scraped data with GfK data. The names were normalised (removed endings like SE or colour). Between 76 % and 86 % of the white goods models in the GfK data were also in the scraped data. The remaining models are not on the market anymore, only sold in physical stores, or in the data set but not recognised by the algorithm. When we took a subset of the models with a cut of after 90 % accumulated market share the figures were 83–93 %.

A higher precision can possibly be obtained by a comparison between sets of performance data, but that has not been tested enough yet to conclude if it works or not.

In all, we estimate that the scraped data were good enough for both market surveillance purposes and policy analyses. Thus, we believe that after this initial test and calibration of the web crawler engine, it really is possible to use only scraped data in the future.

In Table 1[3], a summary of the most important differences between purchased and web scraped data can be found.

As noted in the table, sales data are not found directly on the web sites. However, by using popularity/ranking data obtained from Pricerunner or similar sites, combined with assessments of total market sales obtained from e.g. Statistical Sweden, it is possible to deduce sales of individual models to a reasonable degree of accuracy (Buskirk, 2015).

Examples of information possible to achieve

Since the start of the project in 2015, the web crawler has been scraping data from the Nordic market for all kind of products, either already regulated or to become regulated by eco-design and/or energy labelling. In all there are about 25 product groups that are now monitored on a regularly basis, but so far the analyses have been focussing on a few product groups in order to learn more about what kind of information is possible to get.

Thus, in this section we show a few examples, i.e. different products with different sets of information. In summary, it is the following:

General information:

- Snap shots of the market: how it looks *right now*. Gives an immediate picture of the market.
- Time series, showing the development of the market *over time*. Gives an idea about the trends, e.g. on whether the market moves in the right/wrong direction, if at all.

Note: the graphs shown here refers to market share of models only; no analyses of sales are included.

Indication of compliance or non-compliance with the regulations:

- Compliance with eco-design requirements.
- Compliance with energy labelling requirements:

[2] The data is not publically available as default, but could in principle have to be shared on request.

[3] For whitegoods it is unlikely that a large number of models are only for sale in brick and mortar (or concrete) shops, since the majority of the retailers have a multichannel sales strategy. For more complex products, this might not be the case.

Table 1. Brief comparison between purchased and web scraped data.

Parameter	Purchased data ("GfK"-type)	Web scraped data
Cost	High; pay for each dataset	Medium initial cost, low running cost
Coverage of the market[1]	70–80 %	Basically the same (70–80 %)
Time resolution	Typically, monthly or annual; the higher the resolution, the more expensive	Typically, weekly and up; does not influence the cost
Aggregated or detailed data (with respect to product groups or models)	Typically aggregated; the more detailed, the more expensive	Aggregated or detailed; does not influence the cost.
Sales data	Yes	Can be achieved indirectly; see more in the text

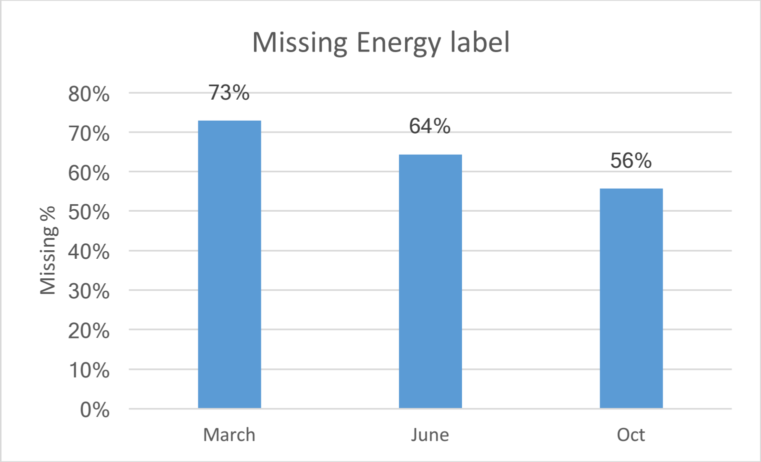


Figure 4. Energy class distribution for Vacuum cleaners, 2016.

- – Labelled at all?
- – Correct label?
- Indications and analyses of possible loopholes used by manufacturers?

Other types of information, useful for policy analysis:

- Adaptation patterns – when does the market start to change and how fast does it move, in relation to when a regulation enters into force.
- Adaptation rates – how fast does the market adapt to new requirements?
- Technology development
- Price development

EXAMPLE 1: VACUUM CLEANERS

In Figure 4 it is shown how the share of non-labelled vacuum cleaners are slowly disappearing from the market, but that many non-labelled models are still on the market despite that the energy label requirement came into force 1st September 2014. Additional information on when the products were put on the market can be used as a guide when deciding whether a formal market surveillance process should be initiated or not. It also gives a possibility to analyse the adaptation rate of vacuum cleaners, and probably insights to why they in this case are slow to adapt.

For the products which are labelled, Figure 5 clearly shows how the market moves to higher energy classes as time goes by; even in the relative short time period of 6–7 months, the share of A-class products increase from 49 % to 60 %. Note again that this refers to the number of models on the market, not the sales.

EXAMPLE 2: WASHING MACHINES

As another example of market transformation, Figure 6 shows how the market for washing machines moves to higher energy classes as time goes by; here it can be seen that the share of A+++-class products increase from 50 % to 60 % in a year.

From a policy perspective, this seems to be a sign of a successful policy. However, since the energy efficiency index (EEI) used to determine the energy class is based on a calculation where the energy use (in principle) is divided by the washing machine capacity, Figure 7 indicates that the increase in the latter in the available models in part may explain why the energy classes becomes better[4]. If combining the two parameters model by model (not done here), it is possible to see if this indeed is true or if it just a coincidence.

[4] The standard deviation is +/- 1,4 kg (at the 68 % level) and +/- 2,8 kg (at the 95 % level) from the average. A t-test was done comparing the Nov 15 and the Oct 16 values, showing no significant change; however, we do believe that as time goes by it will become evident if it is a genuine trend or not.

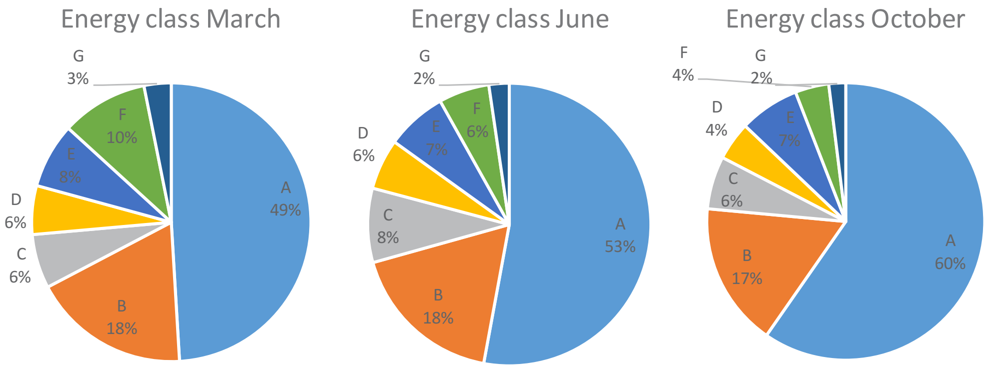


Figure 5. Energy class distribution for vacuum cleaners, 2016.

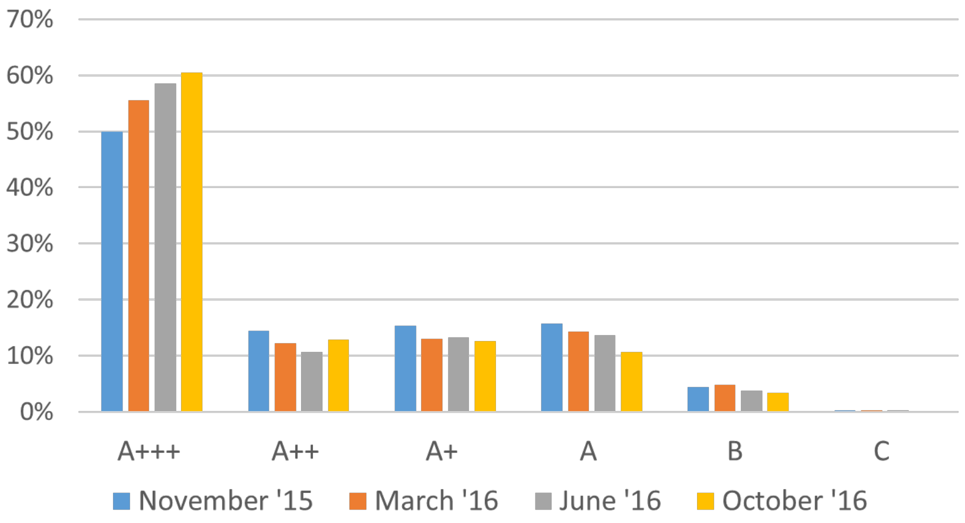


Figure 6. The development of the energy class distribution for washing machines in Sweden, 2016.

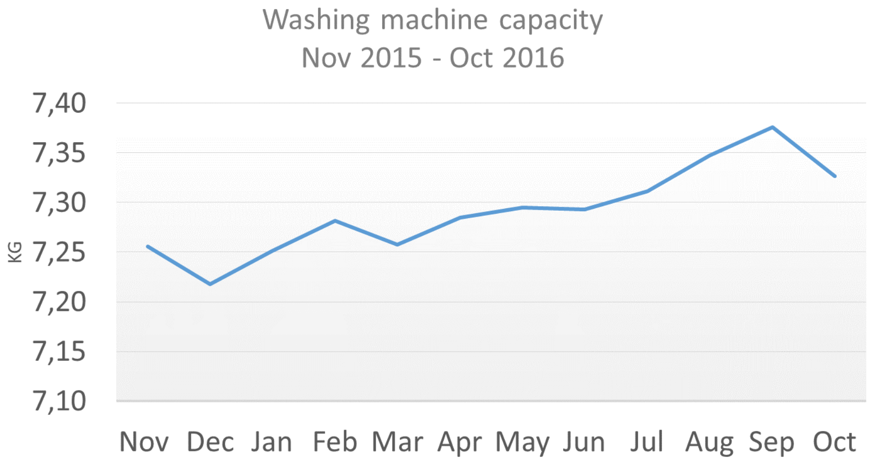


Figure 7. The development of the washing machine capacity for washing machines in Sweden, 2016.

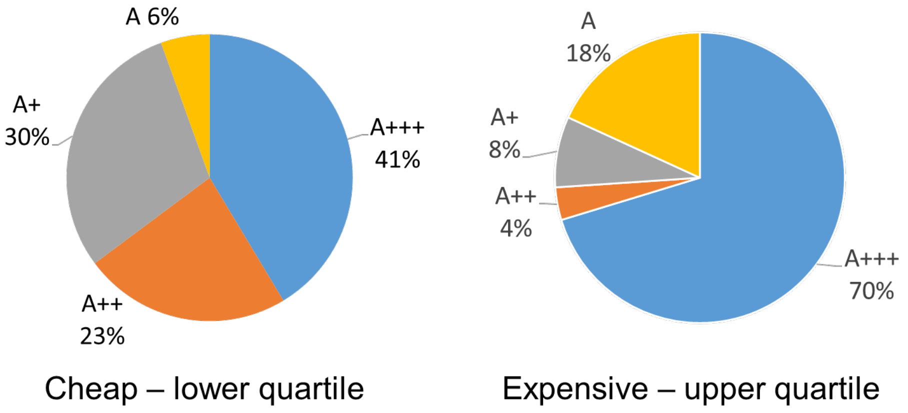


Figure 8. Two different distributions for energy class and purchase price for washing machines in Sweden, 2016.

Finally, Figure 8 shows how it is possible to relate the purchase cost and the energy classes; *on average* a higher energy class is more expensive. However, combining these data with information on life time and annual use, the full LCC can be calculated and compared for different models (not done here).

EXAMPLE 3: COMBINED FRIDGE/FREEZER

In Figure 9[5], the evolution of market share of energy classes for combined fridge/freezers are shown – as another example. However, in this case, it is also shown how a majority of the models sold on the Swedish market are labelled for subtropical or tropical conditions; there is an allowance for higher ambient temperature that affects the energy class and the question is whether this is used as a loophole or if it just happens to be that the same models are sold around the world. Further analyses will be done to check this.

EXAMPLE 4: LIGHTING

The final example concerns lighting, and here we show some other kind of information that is possible to get. In Figure 10, it is shown how models associated with different lighting technologies leaves and enters the market. The example shows that many LEDs leave the market, but also that most of the new models are LEDs. Clearly this is the technology in focus.

As a complement, another perspective on the different technologies is shown in Figure 11: here it is clearly seen how much more efficient LEDs are for the same luminous flux, and, combined with the information in Figure 10, how LEDs are taking over for a flux up to a 1,000 lumens or so.

Another type of information, which concerns the stock of luminaires, is found in Figure 12: here the shape and light base for light sources on the market are shown. The retrofit market will probably dominate for a long time, but when new types of lighting fixtures and systems enters the market, that will be reflected on the share of light sources as well. To get a more complete picture, data for luminaires can be scraped as well.

The final graph in this section focus on the supply of light sources provided by a "old" and "new" company on the market, respectively. In Figure 13 it can be seen how the old company have a mix of all technologies, although the trend moves towards LEDs; going from 17 % (Oct 2015) to 31 % (Oct 2016). The new company, however, is mainly offering LEDs already in 2015 (88 %), and even more in 2016 (92 %). (Note: the incandescent and halogen light sources are mainly for decorative lighting, it turns out from the data; not shown here.)

This clearly shows how new actors can enter a market and immediately go for the state-of-the-art technology; on the other hand, the old companies have to transform their portfolio and the pace at which this can be done will determine whether they can stay at the market or not. Thus, detailed data with a high time resolution can reveal the dynamics of the market, which can be used for studies of innovation rates, learning curves etc.

Conclusions and future work

In this paper, we have tried to show how web scraped data can provide a new data source used for market surveillance, i.e. as part of a tool where different sampling strategies can be used. We have also indicated how the analyses can be used for policy evaluation, either for a deeper understanding of the market dynamics and/or for estimates of energy savings[6] achieved thanks to product policies.

Using web crawler techniques, it is easy to follow the market in *real* time, as well as *over* time. The initial setup of the system has taken a couple of years at a total[7] cost of about 245,000 Euro. Since the running[8] costs are low, there are no real limits to the degree of details or time-resolution that can be used, as long as data can be processed and stored in a systematic way.

The NordCrawl system is still in its infancy and provided it will become an integrated part of the Nordic MS activities, it will take some time to learn how to use its full potential. We do believe, however, that some of the advantages will be:

- Reduced cost for data collection

[5] If a product has multiple climate classes, the climate class that gives the highest correction factor is displayed, according to (EC) No 643/2009 of 22 July 2009 table 6 (http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009R0643&from=EN).

[6] See another paper in the eceee 2017 proceedings, Estimation tool for National effects of MEPS and labelling – version 2.0, 8-238-17 Stengard.

[7] Covering development of the software, initial tests and comparisons with GfK-data, meetings and travel costs etc.

[8] Covering data scraping, data analyses, maintenance etc. At the time of writing, we don't have an exact number of this cost but expect it to be less than 40 000 Euro/yr.

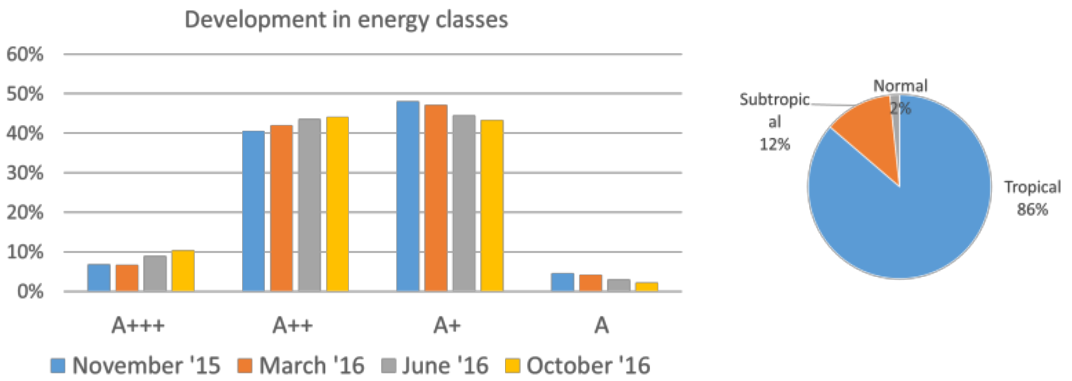


Figure 9. The development of the energy class distribution for Combined fridge/freezer in Sweden, 2016. For October 2016, the distribution of climate classes is included. (The distribution between Nov 15 and Oct 16 is basically the same.)

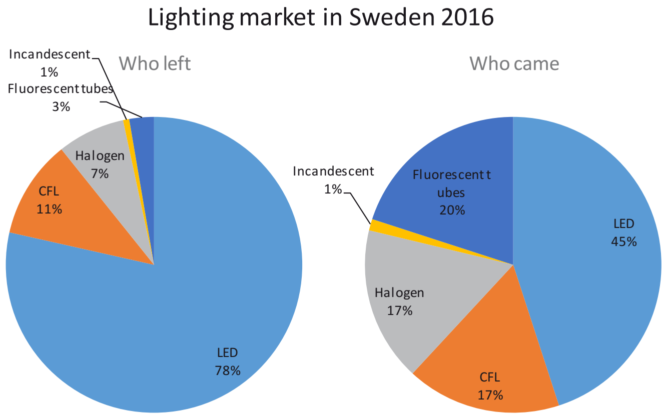


Figure 10. The crawler makes it possible to track which models enters or leaves the market, here grouped by technology.

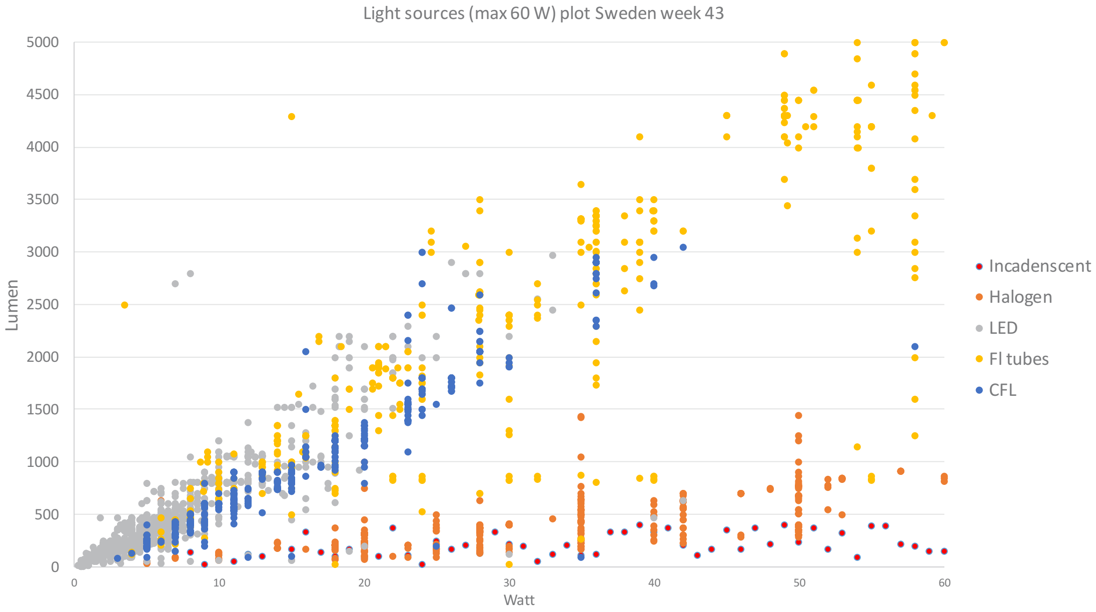


Figure 11. Scatter plot of the Swedish market week 43, 2016.

- Higher coverage and a better representativeness of the market (market picture)
- Data collection in real time
- Improved sampling strategies when choosing products to check, be it random or targeted sampling
- Faster response times for targeting non-compliant products
- Overall improved effectiveness and reduced cost for market surveillance

Right now, we are crawling about 25 products on the market on a weekly basis and at a fairly high degree of details. Time will tell what type of analyses that make sense and what degree of details is required, which will allow the system to be optimised.

References

(Buskirk, 2015) S Touzani and R Van Buskirk. Estimating Sales and Sales Market Share from Sales Rank Data for Consumer Appliance. LBNL-6989, Pre-print version submitted to Quantitative Marketing and Economic, February 2015.

(Nordsyn) See http://www.norden.org/sv/tema/tidigare-teman/tema-2015/green-growth/statsminis-trarnas-aatta-projekt/samarbete-om-groena-teknis-ka-normer-och-standarder/nordsyn-surveillance-cooper-ation-for-green-products (link downloaded 2017-01-22).

(SEA 2017) Report from Magnusson Advokatbyrå (legal firm) 22 December 2016 to the Swedish Energy Agency, and an amending report from NVE Legal department, 25 January 2017.

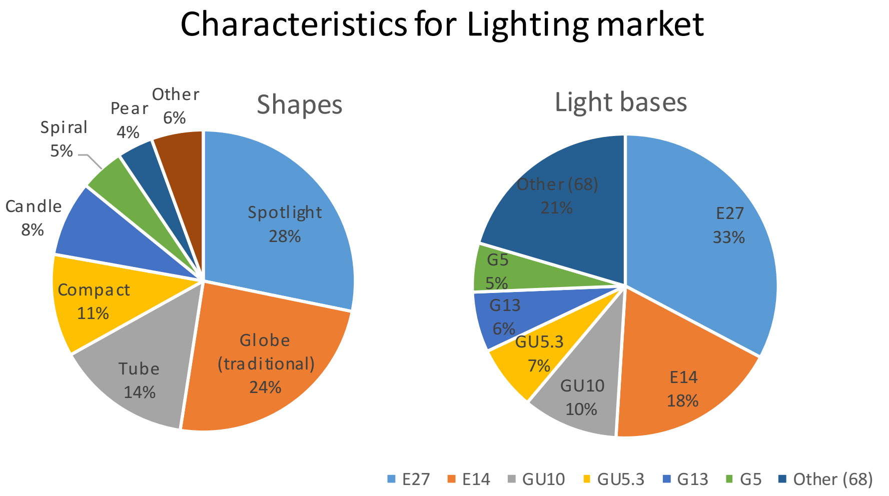


Figure 12. Swedish Lighting market autumn 2016, special features.

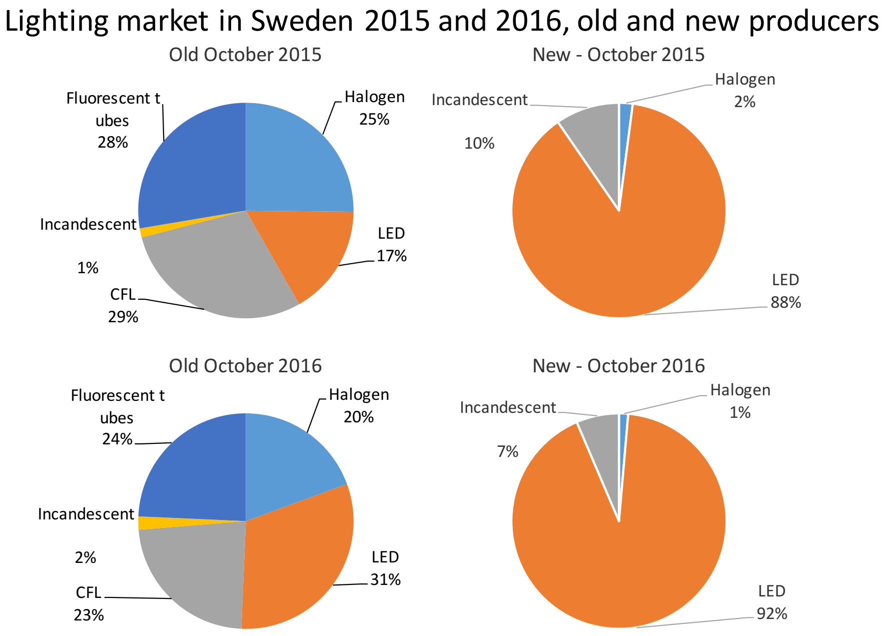


Figure 13. Lighting market analysis, showing what old and new producers offered in Sweden in 2015 and 2016.

(Kemna, 2015) R Kemna and L Wierda, Ecodesign Impact assignment, report to the EC, May 2015: https://ec.europa.eu/energy/sites/ener/files/documents/Ecodesign%20Impacts%20Accounting%20%20-%20final%2020151217.pdf.

Acknowledgements

The authors would like to thank colleagues at the Chinese National Institute of Standardisation (CNIS), the US Department of Energy (DoE) and US Environmental Protection Agency (EPA) for valuable comments.