Impacts of Open Data in Luxembourg and the Greater Region - 2019

Date: 06 May 2019
ISBN: 2-919941-23-2
EAN: 9782919941230

digital luxembourg

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY | LIST

1 CONTENTS

1.1 TABLE OF FIGURES

INTRODUCTION

Three years after its official launch in April 2016, the *data.public.lu* portal has currently more than 800 published datasets and references nearly 110 reuses with the participation of 120 organizations.

After a first experience in 2018, Luxembourg Institute of Science and Technology (LIST) was mandated in 2019 to conduct an evaluation of the impact of Open Data in Luxembourg. In order to better understand its users and stay tuned to their expectations in terms of content and functionality, a satisfaction survey was conducted.

Open Data is grounded on the openness principle, which is designed to minimize the burden on re-users, but it largely deprives data providers of the means to know the re-users, the intensity of the re-use, the modalities of this, as well as the value created from the public assets (because of the free of charge principle induced by the marginal cost). This therefore requires a very broad approach that draws on all available means to capture and analyze the traces of re-use, which implies adopting a methodology taking into account different types of indicators, quantitative or qualitative, at different scales.

This document summarizes the methodological guidelines followed for this study and the main results. It is organized as follows: methodologies followed for the surveys and the analyses of the available logs are explained (part 2); then the main findings from the surveys (part 3), the analysis of logs (part 4), and the insights gained from the first Open Data re-user group (part 5); the document is concluded by the perspectives drawn for the Open Data ecosystem in Luxembourg (part 6).

2 METHODOLOGY TO ASSESS THE IMPACTS OF OPEN DATA IN LUXEMBOURG AND THE GREATER REGION

The methodology of the evaluation tried to combine the different approaches of the state-of-the-art and to adapt them to the profile of Luxembourg, adapting them to the realities of a country to the number of re-users necessarily limited, for a country which does not benefit of the mass effect, taking also into account that the local ecosystem of Open Data is still growing.

2.1 SATISFACTION SURVEY ON DATA.PUBLIC.LU

In order to assess the impact of open-data in Luxembourg, a qualitative study concerning:

- the access of the data
- the content of open-data portal
- the purpose of the use of Open Data portal
- the usage of the data and then the identification of re-users

has been conducted through a general users satisfaction survey during March 2019.

This online questionnaire has been promoted on social networks (mainly *Linkedin* and *Twitter*) and through emailing lists from the 2019 Hackathon and a workshop addressing Open Data for mobility and transport.

The structure of the survey is divided in 4 parts:

1. General survey about data.public.lu (cognition, ergonomics, and motivations for use)
2. Overall questions about the data available on the web site (availability, accuracy, accessibility and quality)
3. Specific focus on observed impacts and topics or format interests
4. Pinpoint highlight about users motivations and context

For this study, we decided to reach as many people as possible with an accessible survey, with around 20 items, and doable in less than 5 minutes without comments. The aim of this choice was to gather enough responses to have a large overview of users of the portal, even the non-profit one.

We gathered 54 responses, from 50 respondents. The aggregated results of the study are presented in this report.

2.2 LOG ANALYSIS

Generally speaking, logs analyses are led following mainly two goals. First, they are used to prevent or detect cybersecurity issues. This objective explains also why they are gathered and stored. Second, they are used to analyze user behavior with the goal to improve the design or the functionalities of a service.

Log analysis make it possible to evaluate at least the audience, and Internet audience is the matter of Web usage mining disciplines. At a slightly higher level of abstraction, they allow to understand usage, i.e. user behavior. For example, *The Analytical Report 8*[1] (p. 7) promotes the analysis of the portals'

[1] https://www.europeandataportal.eu/sites/default/files/edp_analyticalreport_n8.pdf

log files as a way to do so. At an even higher level of abstraction, one can study some of the impacts, or at least the conditions of the impact.

Depending on the service considered, the format chosen by administrators, log files store different kinds of information. Most frequent fields are date, IP address, kind of action and time required to answer.

Logs are useful to gain insights on the data lifecycle after their release. More accurately, one can state that they allow catching the impact while it is happening from the perspective of the access to the data. Another advantage is that this method is comprehensive, except any technical issue, as all the interactions are recorded. A drawback is that one has only a view from the perspective of the access, but the gaps from the perspective of the re-users, the re-uses and what the final users are doing with them are important. For these, we have only partial information, allowing to build hypotheses that shall be filled and confirmed by other means. Therefore, this analysis is led in conjunction with other tools, i.e. surveys. However, rich and deep are the metrics available; they need to be combined with different data. That is why, for the different datasets, a man-made post-processing labelling has been led, intending especially to label the datasets with the thirteen themes used by the European Data Portal (EDP). If it is a very slow and time-consuming approach, and not easily automated, but it is fruitful to assess the areas the most prone to generate impacts from Open Data.

One purpose of this study is to test if different kinds of logs allow emitting and checking more or less precise hypotheses on the understanding of re-users behavior and on the impacts of the Open. Data.

Access and downloads logs also allow assessing the audience of public data through their categories and their formats.

Query (or search) logs, combined with entry pages and even referrers allow to make assumptions on what is sought by re-users and to assess if they find this easily.

Two specific issues are dealt with in the fourth chapter of this report: are logs analysis methodologies suitable to analyze the impacts of one kind of data (i.e statistical data) and to monitor the impact of the initiatives aiming to trigger the economic potential of their data (i.e. the participation to the hackathon Game of Code).

3 SATISFACTION SURVEY RESULTS

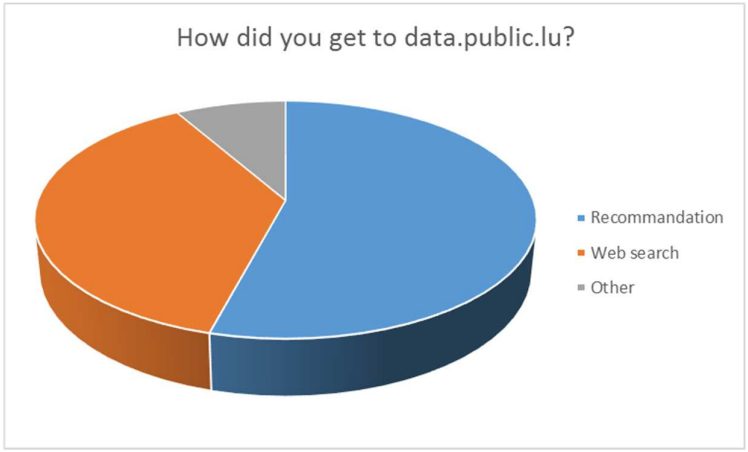
The satisfaction survey about data.public.lu has been conducted through *limesurvey* and communicated on *Linkedin*, *Twitter* and contacts list from Open Data related events.

We gathered 54 responses, 18 of which were completed.

The main findings are detailed in the following sections.

3.1 GENERAL SURVEY ABOUT DATA.PUBLIC.LU

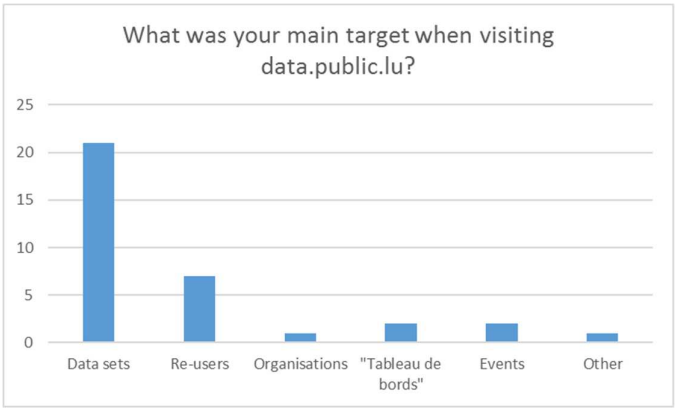
The first part of the questionnaire concerns data.public.lu.



It appears clearly that the major part of visitors are aware of the existence of this platform, either from adviser (LIST, Ponts & Chaussées and cadastre administration were quoted) or social media communications (LinkedIn). The web research part represents google searchs.

Figure 1 - Accessibility

The visitors of data.public.lu are mainly interested in finding available datasets.



Datasets answers include:

- "datasets for big data, similar to Kaggle.com"
- "Raster data, data about Luxembourg"
- "Roadworks aso (Ponts & Chaussées)"
- "DCAT AP related data"
- "Wms tiles. Lidar point. Dem. Buildings. Roads. Addresses"

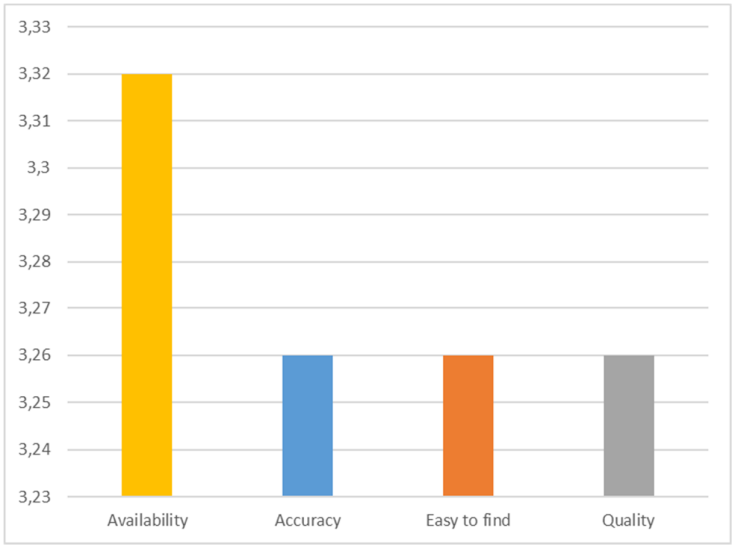
Figure 2 - Target

By targeting re-users, respondents explain:

- "Find reusable software applied on datasets similar to Kaggle.com"
- "Try to get new uses/derivations to ameliorate my own dataset"
- "Concrete Projects"

One respondent admitted just wanting to explore the portal and its functions

3.2 FOCUS ON THE DATA



Concerning the data, the respondents feeling is quite mitigated, but still promising. Availability get the best quote (3.32/5 mean value, and 1.03 standard deviation) but accuracy, accessibility and quality remains close (3.26 mean value, and standard deviation around 1)

Figure 3 - Data

The detailed results are represented in the following graph:

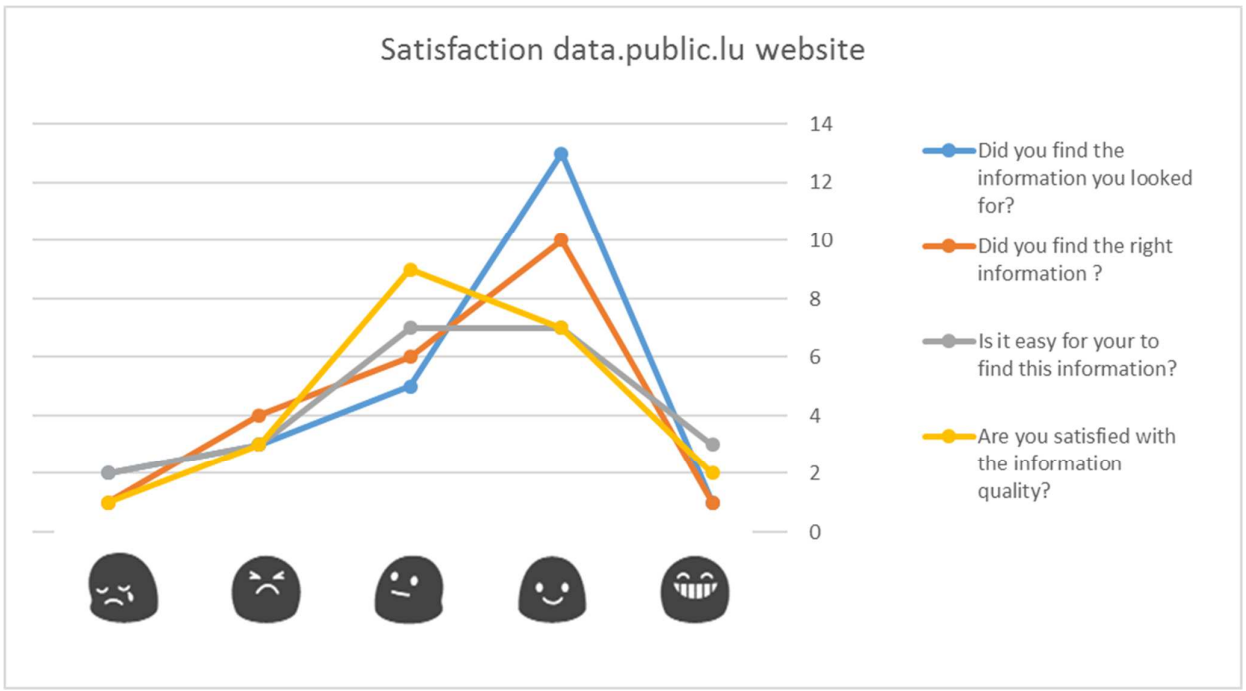


Figure 4 - Satisfaction about data

Remarks:

- ”Knowing Open Data Portals of Germany and Austria, the Luxembourgish one is a very painful experience. Documentation is so bad, the datasets are broken and incomplete (Inspire) and this are datasets coming from the administration. A linkage in the geoportail to the currently used dataset (if applicable) would be a great idea”
- "Datasets remain weak"

We asked the respondents if **some item disturbed** their visit. The answers are:

- “Documentation of the datasets, quick stating of the metadata of the datasets”
- “Few data available that are truly reusable for innovative services and new products”
- “Sometime the search didn’t always return the datasets I was looking for”

We also ask them if there was **one thing to change**:

- “Ameliorate the datasets and uniformize them”
- “Dataset count and publisher count in the first page”
- “Visibility and richness of content"
- “There are a few datasets that don't have a specified license, in the context of an Open Data portal, this is rather hilarious. Maybe discourage uploaders from choosing a license?”
- “You must have real time Open Data at a rate that allows real re-use”
- “Add advanced search”

3.3 FOCUS ON THE ADDRESSED DOMAINS

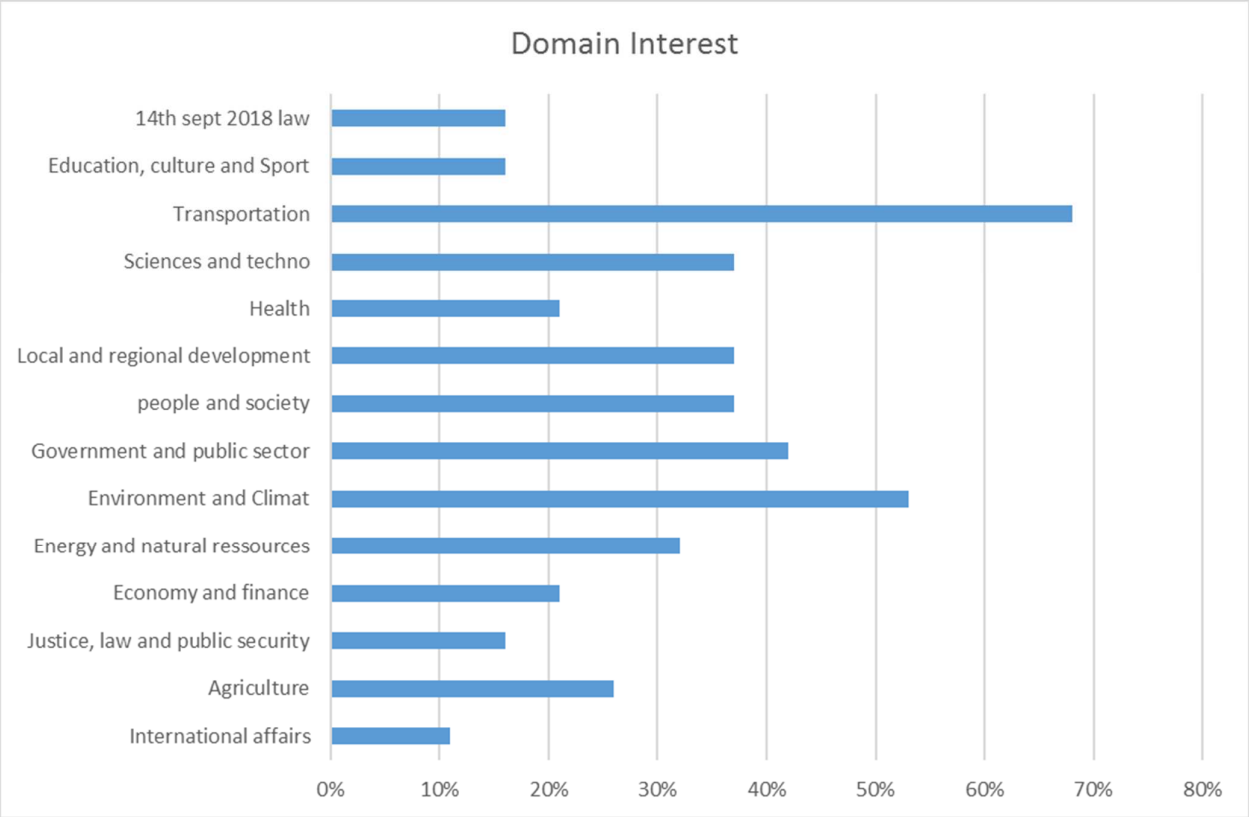


Figure 5 - Domains

The major topic of interest is Transportation, but this information should be considered knowing that we contacted people that attend a workshop on Open Data in transportation.

Nevertheless, the three main interesting domains are (i) Transportation, (ii) Environment and Climate, and (iii)Government and Public Sector.

Respondents had the possibility to express a focused interest for each domain. The list above describe their comments:

- For Justice, law and security: Neighbourhood crime rate
- For Economy and finance: Socio-Economic indicators
- For Energy and natural resources: distribution of natural resources
- For government and public sector + Local and regional development: Future building projects and urban extensions
- For People and society: RNPP statistics (Registre National des Personnes Physiques)
- For Health: Localisation of doctors, pharmacies…
- For Transportation: Bus stops, bus frequency…, ...

The **missing data requested** by the visitors are:

- “Larger datasets in easy-to-use formats for analysis and machine learning, e.g.: CSV, JSON, XML”
- "Socio-ecomic datasets"
- "Registre du commerce"
- "Juridique : jurisprudence, doctrine
- "Données historiques : vieilles données du recensements par exemple ou autres. données météo historiques manquantes aussi"
- "Toute sorte des données transport logistique, mobilité, transport en temps réel réutilisables pour des développeurs des 0uvelles services et produits. "
- "Taux de criminalité, Projets futurs de construction et extension urbaine"
- "Small area population data"

3.4 Impact of Open Data

In this section, visitors had the possibility to express their vision of the impact of data.public.lu data .

All impact domains are validated, with a larger social impact.

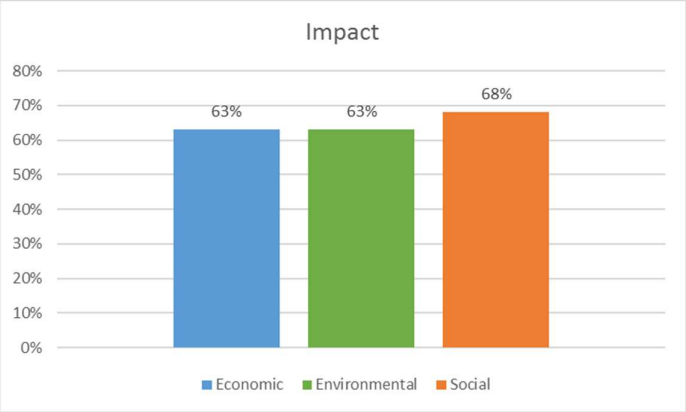


Figure 6 - Impact

The explanation of economic impact is the business opportunity for ITC and data science market, and jobs or start-ups creation possibility. It also allows more efficient policymaking and strategy definition. On the economic side, land rezoning and Urban planning also create business impact.

On the environmental side, the data provided by the portal can be used to provide scenario and predictive modelling. ie flood risk.

Social impacts are the more consistent: respondents consider that the portal facilitate citizen science initiatives and gives a better understanding of population need, demand and use.

3.5 TECHNICAL FOCUS:

The claimed format requested by respondents are:

- JSON
- CSV
- XML
- LUREF and shapefiles for GIS use
- plain text (avoid pdf or binary files)
- Any machine-readable, open and/or widely used format is fine.
- Real time

The others Open Data platforms used by respondents are:

- Kaggle.com,
- Google Datasets
- Austria Open Data portal,
- USGS data portal
- UK Open Data portal: data.gov.uk
- European Open Data portal
- data.overheid.nl

Satisfaction results in a nutshell:

- *More communication and advertising about the portal is needed.*
- *Finding datasets is the main goal of visitors*
- *Datasets should be expended and improved (real time, documentation, Inspire data-sets, harmonisation of data-sets)*
- *An advanced search tool is requested*
- *All the domains are validated (with improvement ideas)*
- *The socio-economic impact of data.public.lu is real*

3.6 RE-USERS IDENTIFICATION:

3.6.1 By context

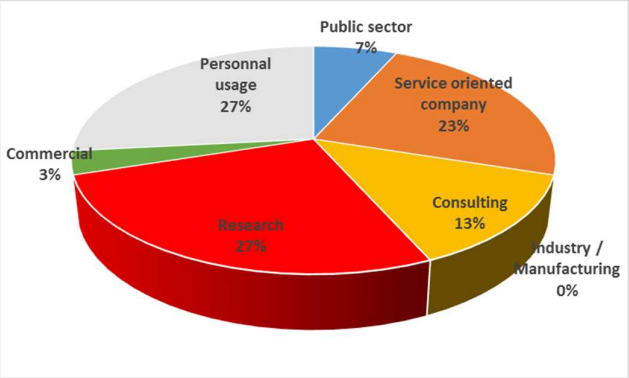


Figure 7 - Context

Re-users are mainly professional users, from private sector for 49%.

Research represent 27% of the re-users and public sector only 7%.

Nevertheless, 27% of respondents are using the portal for personal reason.

3.6.2 By size of companies

The major part of private sector re-users are working in Medium or large enterprises.

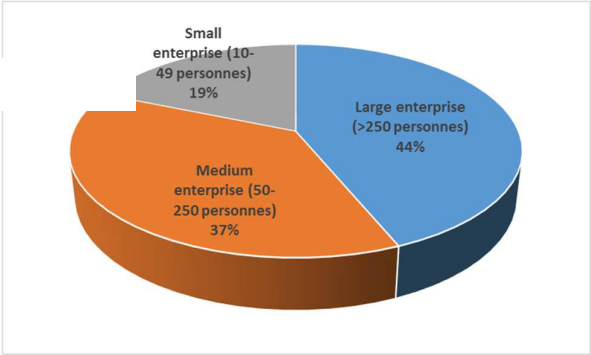


Figure 8 - Size

3.6.3 By country

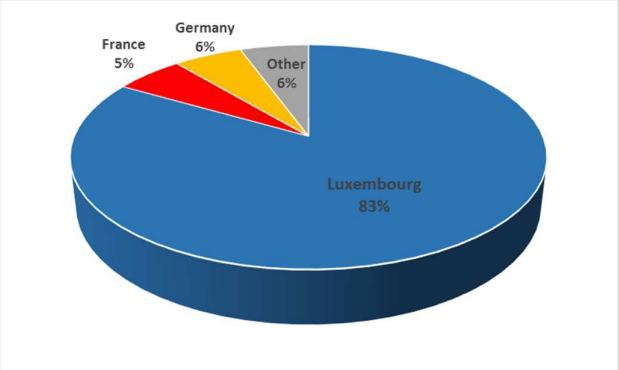


Figure 9 - Location

Since our study concerns the Luxembourgish Open Data portal, the large majority of respondents works in Luxembourg or neighbouring country.

Re-users identification results in a nutshell:

- *Re-users are willing the portal to improve*
- *Re-users are 1/5 research, ½ private companies, and 1/5 for personal use.*
- *75% of re-users are working in medium or large companies.*

4 LOG ANALYSIS

As defined in the methodological guidelines, this section intends to extract the information related to the audience of the platform (mainly through the pages viewed by the users as well as the recorded downloads), then to assess the potential contribution brought by the query (or search) logs to assess what the re-users intend to find on the platform. Compared with the previous report, this document is committing to analyze the advantages (and the limits) of the provenance of the users, which can be summarized in this way: what can one learn from the visitors given their provenance. Again, compared with the previous edition, this report is focusing on two specific issues, related to the suitability of log analysis methodologies to monitor the impacts of statistical data and the impacts of a hackathon organized in 2018.

To lead this analysis, different sets of data were provided: (i) logs (both access and queries) of the national Open Data portal for one year, (ii) same data but only for the four days corresponding to the hackathon in March 2018, (iii) same data for one year, but only the datasets provided by the national agency for statistics, (iv) data provided by the STATEC: one month of the query logs captured on its platform.

4.1 MAIN INDICATORS

The following paragraphs are defining the main indicators available for the log analysis.

Visit – or session in other tools - is the basic metric[2]. Following the settings of the tool used to gather the Open Data platform's logs, a **visit** is recorded as such any time a **visitor** is connecting to any page of the Website. If a visitor is interacting with the platform more than thirty minutes after the last page viewed, it will be recorded as a new visit.

'**Pageview'** metric is the number of times a given page has been loaded during a given time period. Among the available metrics, it is providing the finest granularity. In Web Analytics literature, it is equivalent to **hits**. **Unique pageview**[3] is equal to the number of sessions that have included at least one visit of the considered page. Given the definition of 'visit' it is not possible to establish a direct link and to consider that 'unique pageviews' are triggered by unique visitors.

One important metric of web analytics is "**unique visitors**", or unique users. It aims at gathering the unduplicated number of visitors a for a given period of time. The concept of unique visitor is facing several limitations. From a technical standpoint, it is a complex figure defined as the "number of unique combinations of an IP address plus a further identifier". It is rising technical issues, as it is

[2] https://glossary.matomo.org/#metricsV

[3] « The number of visits that included this page. If a page was viewed multiple times during one visit, it is only counted once. »

estimated from a series of other indicators, e.g. user argents or cookies, but no statistics are available on the average weight of each. These methods imply large uncertainty, as cookies are often overcome in current browsers. Thus, it is ambiguous, as the figures built are counting devices, not people, whereas terms such as visitors or users tend to refer to human beings[4]. Methodological constraints must also be taken into account: for each title page or for each URL, there is an annual total, but this is based on a daily basis. So, the annual number would be different if the system was weekly or monthly. In other words, this indicator should be considered as the **annual total of unique daily visitors**. Despite its gaps, this sum provides a useful ground for comparisons.

Users are a very close metric taking into account the visits done by registered visitors. It may be considered as a metric providing more insights on how are re-users accessing a given resource. For these reasons, this metric is also minoring some of the drawbacks mentioned for visitors, albeit not completely. Registered users tend to be a very specific kind of visitors, not representative at all of the visitors at large.

For each visit, the consecutive **actions** are recorded. The recorded interactions are pertaining to clicks leading to the start of a download, the use of the platform's search engine, the visit of another page (on the platform itself or on an external Website).

Bounce rate is the percentage of entries on a page of a Website that are not followed by another page view. Averaged at the scale of the Website, this metric is helping to balance the figures provided by the 'Visits' metrics, and thus is useful to get a more subtle picture of the audience. It is closely related to the **exit rate**, i.e. the percentage of visits ending after the consultation of a Website page. For example, a kind of ideal visit path would be the case of an user landing on the main page of the portal, where he would use the search engine, then would view a page presenting data relevant for his requests, finishing finally his visit on the download page that would be the exit page. Bounce and exit rates are not equivalent as, for example, a high exit rate may be only the consequence of a successful research, especially on the data pages, while a high rate could be a good indication of high accidental visits.

4.2 MAIN FIGURES ON AUDIENCE

To understand how the frequency of access and use may feed the analysis of final impacts, one has to answer several questions: **is the platform used? What can one learn concerning the frequency of use? From the methodological standpoint, what are the relevant metrics? the benefits and the limits of using logs data to do so?**

[4] http://www.ifabc.org/pages/news/metricchanges.html ; https://web.archive.org/web/20100911060028/http://www.webanalytics.com.br/Recursos/Ingles/IFABC_Web_Standards.pdf

The first indicator considered is the number of **pages viewed**. From December 2017 to December 2018, the added figure is **562 440, an average of 1 540 pages viewed per day, and an increase of 40% since 2017**. Over the same period, the number of unique page views is 316 632 but it quite a weak measure, as these unique page views are computed for specific URL and adding the results for all the URL does not provide a figure of different visits, as very large overlaps may be expected. The number of **pages viewed is providing an indicator of the audience or popularity of the platform** and shall be compared with other indicators computed for the platform, with other economic or demographic indicators for Luxembourg. For example, they may be compared with the indicators computed in the 2018 ex-ante analysis of the economic impacts of Open Data in Luxembourg. The reports made for the European Data Portal suggest that people engaged in knowledge intensive activities are the part of the population the most prone to engage with Open Data, estimated at 132 500 in 2016 in Luxembourg by the national statistical agency. Moreover, the 2018 estimated the number of jobs based on Open Data re-use at 100 in 2016 (and 400 indirect jobs) with a forecast of 150 in 2020 (with 550 indirect jobs). **Based on these assumptions, one finds a potential ratio of 4 pages viewed per person employed in these activities**.

A more detailed analysis of the URL accesses is allowing to **categorize the 500 most popular resources for which detailed information is available**, as the platform is gathering different kind of resources beyond datasets. Results are very similar with those found for 2017, as around 10% of the resources could not be categorized, and **datasets increase their share from 75% to 80% of the pages viewed**. However, (most popular) pages about re-uses are gathering around 40 000 views. Even if one may consider than infomediaries are more prone to access data and general public to focus on re-uses, the available data do not allow to confirm it, hence the need to cross these insights with other methodologies.

4.2.1 Narrowing the estimates

These figures shall be considered as a maximum, as several factors may decrease the actual figure.

The number of views is influenced by the architecture of a website: the more modular it is, the more it will overestimate the number of pages viewed for only one interaction. Accidental accesses may also exaggerate the actual audience, requiring post-processing, for example through the analysis of the bounces. The problem is identical with the impossibility to make a strong and clear difference between general public and professionals. Conversely, this method has also an underestimating bias, as it is only considering a sub-set of the re-uses and their final impacts.

To mitigate this issue, several indicators may be combined. A ratio between the page views and the unique page views (1.7, stable compared with 2017) was computed, meaning that **in average, for 1 unique page view, the visitor is coming back to view it again 1.7 times**. Again, the link with the access is weak: if one visitor is downloading the data and leaving the platform, the link is broken and it is not possible to analyze accurately whether there is an use, and if it is intensive.

The same limits stand for the **unique visitors** - or more accurately for the unique devices-, giving 200 649 and **a ratio of 2.8 compared with page views** for the downloads. Downloads and bounce rate are showing respectively a large increase and a stable ratio since the analysis of the data available for 2017.

The most interesting insight considering the precautions to use these indicators is that the data have been gathered for the same period and following the same methodology and tool for two years. It means that **whatever the limits and the uncertainties they are bearing, they show a clear increase of the interactions, at variable but high growth rate.**

Indicator	**Value**
Added number of page views	562 440
Evolution since 2017	+40.23%
Number of unique page views	316 632
Evolution since 2017	+37.44%
Ratio (page views / unique page views)	1,7
Annual total of unique daily visitors	200 649
Evolution since 2017	+21.3%
Average number of unique daily visitor	550
Ratio page views / unique visitors	2,8
Number of downloads	17 535
Evolution since 2017	+141.93%
Ratio page views / bounces	33.3

Table 1 - *Main metrics of use*

Indicator	Value
Ratio page views / population employed in knowledge intensive activities	4
Ratio annual sum of daily visitors / population employed in knowledge intensive activities	2.38

Table 2 – comparison with statistical indicators

4.3 GEOGRAPHIC PROVENANCE OF THE VISITORS

The possibility to analyze the geographic provenance might be a fruitful source, at least to identify where the impacts are realized. The file region is gathering data for only 56 479 visits – **still an increase of almost 65%** - but less than 10% are actually linked to a specific geographic origin. Unfortunately, the module catching the geographic provenance has met issues during this year, as almost zero traffic is recorded for Luxembourg, whereas the different entities dividing the country were ranked at the third place in 2017.

Using the few data usable, and distinguishing visits and actions, the ratios between these two indicators show that few re-users are triggering a lot of actions: maybe some unique re-users proceeded to a systematic (and probably automatic) download of a large part of available data. This ratio is strengthening the ideas of a regular analysis by search engines' robots, or of regular visits to download the data or to update the previously downloaded data. **If the average ratio decreased compared with 2017 (from 75 to 17), this is obviously due to technical issues**.

4.4 QUERY LOGS ANALYSIS

Query log analysis is a complementary analysis providing different insights. If the pages viewed are showing how the visitors are interacting with what the platform is providing, **query logs are showing what the users intend to find on the platform, what are their needs, and (partially) what are their intentions**. It is possible to link the frequency of a given term sought with the popularity of a category of datasets, and so identify the most suitable ones to generate impacts.

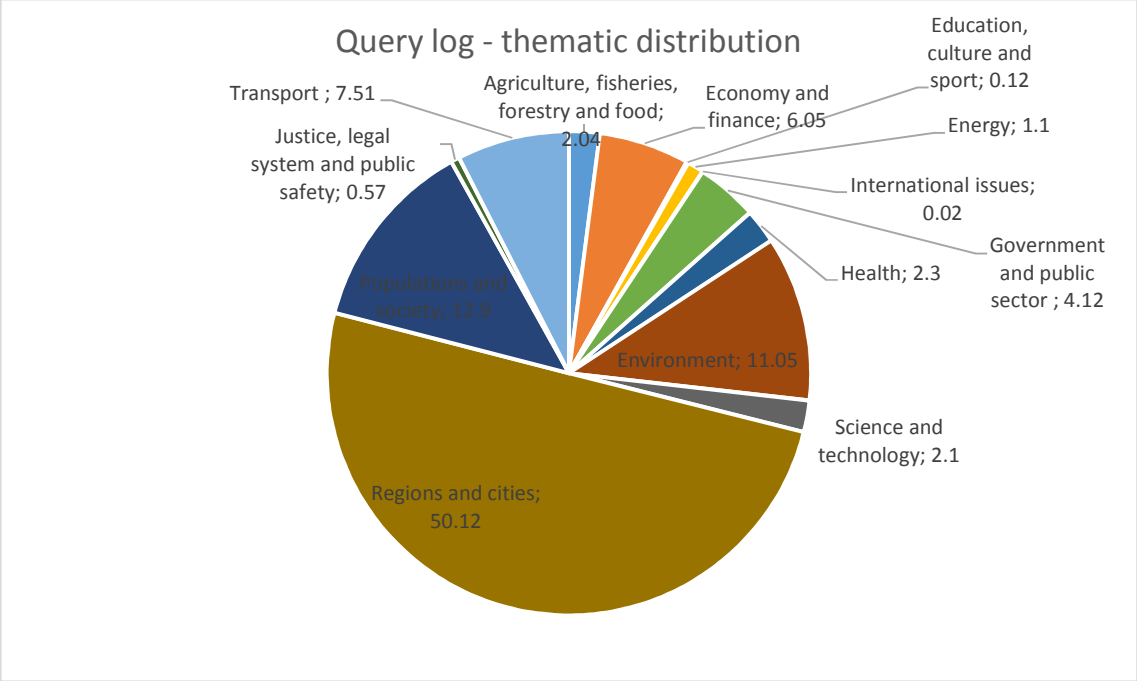


Figure 10 - Query logs – thematic distribution

The datasets have been manually labelled depending on their categories, output is displayed in the graph above. They are very **similar to the results found for 2017: the ranking of categories is not modified**, the trends are even strengthened. For example, the **overwhelming importance of the category region and cities** (mainly due to the labelling of geographic information under this label) is confirmed, confirming also the importance of **geographic information among the high value datasets**. Even the first most sought terms remain in the same order than in 2017, the first one being "ortho", referring to the aerial photography of the country.

From the linguistic standpoint, the majority of data being in French, it is not surprising that most of the queries are in this language. Even, the share of French in query terms is higher than the number of datasets published in French. English is emerging for only 6 terms, on very technical issues. German is counting only three terms, e.g.: "wasser", for water.

4.5 Analysis of referrers

In web analytics, given a website (and a page of a website) a referrer is the website visited directly before. This is a new part of the analysis that was not led in the previous report. Globally, if we put aside the 'other keywords' issue, this view is partially escaping the granularity problem. It is less harmful, while still existing (as we cannot link a referrer to a set of specific accessed datasets).

An analysis of the referrers may contribute to the impact analysis in several ways. In order to generate economic impacts, visitors and especially potential re-users, have to be aware of the existence of the platform and its content**. Referrers are providing a measure of the traffic intensity from different kind of sources and especially different URLs, and so is informing on how and how much external**

channels are improving the visibility of the platform to a public suitable to generate a higher economic activity. This is related to the upstream activities prior to the realization of the impact. It is also possible to leverage it for the downstream analysis, at the very moment of the impact realization, through the insights gained on the re-use of the data themselves, e.g. when and end-user interacting with an application re-using the data is downloading these data.

More particularly, they provide some information on **the deepness of users involvement (and what is the influence of their provenance on it), on search engine visibility (keywords), allow to make assumption on visitors intentions (keywords), what kinds of data the visitors are seeking for (keywords, beyond and in addition to the internal search engine data), identify the citations from other part of the Internet through URL (Websites & social media), make assumption on browsing patterns (URLS through bookmarking) ; make (even weak) assumptions on who are the visitors (based on URLs)**.

For the use of referrers as part of a steering tool, these basic indicators may be combined to answer some questions, such as the identification of the sources bringing the largest flows of visitors and if they are or not contributing to long-term engagement with the platform. They may also be used to identify and understand what are the weaknesses related to bad-represented sources for which a higher share would be expected. In turn, it will feed different strategies aiming to improve the visibility (e.g.: SEO optimization).

It is also useful to complement other metrics, like pages viewed, adding one condition, i.e. knowing from knowing from where are the visitors coming from, or following which search query and so with which intention.

In spite of its advantages, this view is still facing some limitations. This metric faces an important limit, as for around two thirds of the visits, the relevant information about keyword is not communicated to the platform, following the policy of several search engines or Internet browsers (e.g.: Firefox), harming the analysis potential of this metric. It shall be taken into account for assessing the representativeness of the following analyses. The analytic potential is also hindered as the statistical view is missing the fine combination of each page related to *n* referrer(s).

The table below is summarizing the main insights that can be extracted from these data:

Metric	Value
Percentage of known referrers	31.2%
Average ratio actions / visits	2.98
Average ratio visits / unique visitors	1.03

Table 3- general figures about referrers

Besides the available data, a handmade pre-processing was led to label the referrers:

- In six categories depending on the kind of referrer;
- The owner of the resource (public or private sector);
- The category of the resource, following the domain.

Combining the available indicators and metrics, it is possible to make cross analyses and so to extract information on the kinds of referrers existing, the repartition among these sources, thus the channels used to access the Open Data platform. The table below displays the results of the categorization of the referrers according to their kind.

Kind	**Percentage**
Application	1%
Intranet	1,3%
Keyword	3,6%
Search engine	10%
Social network	11,6%
Website	72,5%

Table 4 – Kinds of referrers

Websites are representing the majority of the known referrers, covering actually a large variety of topics (see categorization below).

Much behind websites, **social networks are representing 11.6% of the known connections**. The relative shares of the social networks on the global Internet are not respected. Facebook, and at another scale Reddit, are in a minority**. More than two thirds of the connections registered from a social Network are coming from Twitter, and less than one quarter are coming from LinkedIn**, which is confirming the importance of professional interests as a factor conveying the streams of visitors.

Search engines (excepted keywords) share a similar weight with social networks. Although they are open to everyone, one can expect they are more prone to attract professional or advanced re-users. From the platform's perspective, they are contributing to the visibility of the datasets. The table below is summarizing the results for a subset, i.e. the search engines specialized in data provision. By far, the **national Geoportal is the most represented**, which can be partially explained by the organic links between the two platforms, but also by the importance of geographic information, as a high value dataset. **Google Dataset Search**, released in the last quarter of 2018, brought some traffic to the platform, representing 15% of the visits recorded for this subset. However, the ratio of actions per visit as well as the percentage of bounces is pretty important given the figure of visits. This could plead for a larger rate of exploratory visits and is requiring to monitor the evolution of this source over time: it is only reflecting a preliminary curiosity or answering a true need? Thus to state if this kind of service is bringing new re-users on a longer term. The European Data Portal is also bringing some traffic to the platform in similar proportion albeit for a whole year.

	Visits[5]	Actions	Maximum actions in one visit	Bounces	Unique visitors (daily sum)	Users (daily sum)
Géoportail	69%	2893	72	336	686	14
Google Dataset Search	15%	455	68	107	170	0
www.europeandataportal.eu	11%	618	148	61	137	2
http://data.europa.eu	1%	142	56	5	19	2
http://inspire-geoportal.ec.europa.eu	2%	40	10	7	13	2

Table 5 - Data search engines

In spite of its low share, **the referrer 'Intranet' is interesting because it is allowing to express an assumption on a professional use**, as the data might be referenced for the internal knowledge of these organizations (from public or private visitors), even if this hypothesis should be confirmed through a research of patterns among the pages actually seen. An interesting example tending to confirm this assumption is the case of Luxmaco[6]. It is not an Intranet per se, but it is very close to the way how Intranet may work from the Open Data platform's perspective. This website is a forum gathering expert information for professional stakeholders of the electricity market in Luxembourg. Most of the content is private, apart from the root page, opened to the general public. The connection from this URL are thus most probably due to expert re-users.

The known keywords represent a very limited part of the traffic, but shall be put in line with the technical limitations explained above, and the actual traffic brought by these keywords shall actually be considered much more important. The **most popular keyword is 'société nationale de circulation automobile' at the 49th rank of the referrers**. This is a public organization handling for example the driving licenses or the license plates. If we ignore the corresponding (data) pages reached through this referrer, it is nevertheless possible to examine the data provided by this organization to the national Open Data platform, related to its activities, provided at the statistical level (e.g.: "Permis à points - pertes de points par type d'infraction 2018"), and falling under the category 'Transport' of the European Data Platform's typology**. The first keywords specific to the platform itself, 'Open Data Luxembourg', is emerging only at the 64th rank**.

Each referrer was categorized as pertaining to public or private body, when it was possible to ascertain it during the manual labelling stage. **60% of the referrers associated with a label are related to a public organization.** This is confirmed with an overwhelming preponderance for the visits and the actions, with a share of **80% of the visits**. Even if one has to take into account some potential bias, first and foremost the small amount of data actually available, it remains that public sector's owned websites are generating a large source of traffic to the platform. As a consequence, the lessons learned from these public websites should be analyzed to ensure a larger area of impact in private sector.

[5] Share among data search engines (rounded down to the nearest unit).

[6] https://luxmaco.vbulletin.net/luxmacoforum

Another question arising, issued from the fact that public sector is recognized as the first re-user of its data, **is whether the visitors coming from public owned resources are also working for the public sector**. With the figures available, we are not able to distinguish who are the users actually brought by these websites. It would be possible to get a narrower estimate (albeit not comprehensive) by analyzing the connections and the IP. In complement, co-analyzing the IP and the pages accessed would be relevant to identify if there are noteworthy patterns contributing to answer this question, and the question of the share between occasional and professional re-users.

Moreover, the referrers were labelled following the thirteen categories used on the European Data Portal[7], when it has been possible to assess the category.

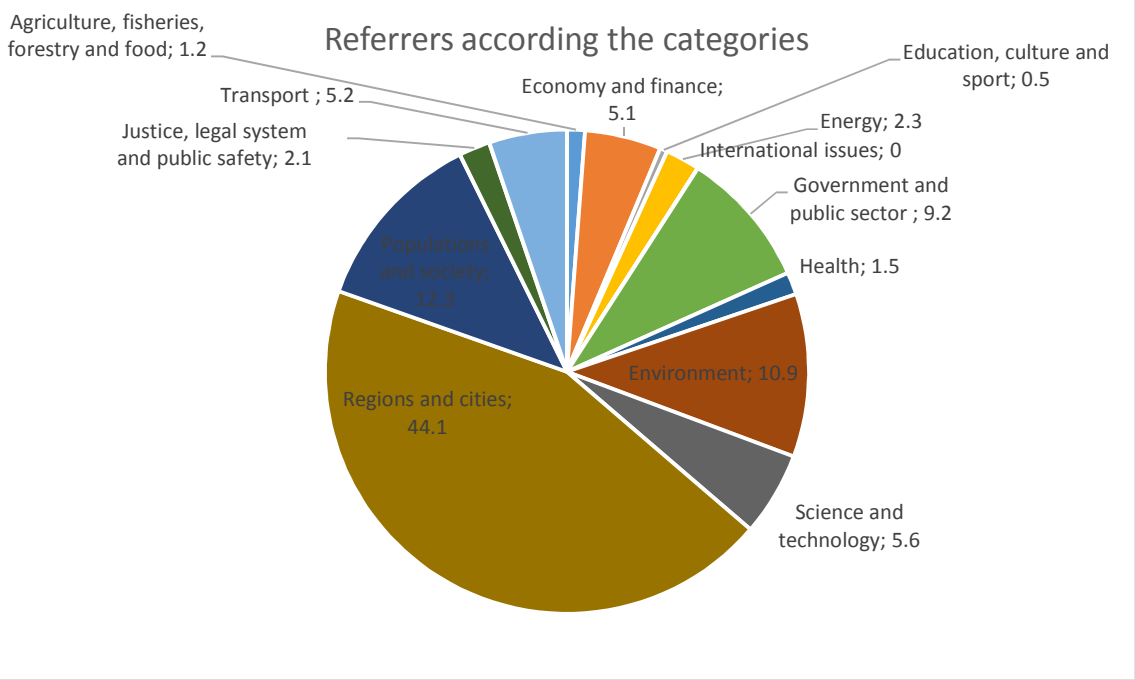


Figure 11- Referrers according to their categories

As they are, these data cannot bring a definitive conclusion to the question of the profile of re*users, if the data are attracting a majority of professional or general public people, or if there is a balanced share (but this would be consistent with the answers to the survey). **Pertaining to the category "Economy and finance", the prevalence of the interest for statistical data may be confirmed** by the high rank reached by external websites providing or dealing with this kind of data, one of the most popular being for example the website of Adem, the national unemployment agency. For the category **Education**, we may only notice some connections from the **University of Luxembourg**, especially from the mail server, and from the **national library website**. This is also relevant for other categories, especially "Environment" and "Regions and cities". The higher share of the "category Science and technology" is partially linked with the search engines dedicated to data, the social networks attached to this category, and .the **code repositories** such as GitHub (23th rank). R-bloggers is also providing

[7] https://www.europeandataportal.eu/data/en/group

some flow of re-users and is illustrating some potential data re-uses. The **category 'Transports' is among the underrepresented categories**, in spite of its identification as high value dataset. This would be consistent with the idea that end-users are not directly accessing these data, but are rather consuming them through the re-uses created by the intermediaries actors.

At a lower granularity level, one possibility is to identify a **traffic from the website of specific companies, while still not being sure that the identified traffic is actually the trace of the visits made by the employees of these companies, but there a good likelihood to be so**. Indeed, the available logs contain the mentions of the websites of companies. At least, it is feasible to make a list of companies and to explore this through other means, e.g. by contacting them, while respecting the privacy concerns. For example, among the companies identified, several are producing services for emergency situations, devices and services for disabled people

Connections from Luxembourgish newspapers are counting for around 500 unique visitors, a modest figure that shall be shaded by the limited number of known referrers. The main newspapers and information websites of the country are represented, e.g. RTL, Luxtimes, Wort or Paperjam. Two large kinds of audience may be brought by these websites. One is consisting in the people reading the articles which are quoting the platform and the data it is hosting. In this meaning, a direction to explore could be the role of the platform as a trusted third-party committed to provide faithful data. This would be a kind of impact a bit different from the others more focused on tangible impacts such as applications or services, but it would be in line with the current concerns on fake news. The second stream of possible visitors may comprise journalists or data journalists[8], who would be a category of re-users to be more specifically engaged to increase the impacts of Open Data. Even if this idea is supported by at least one server dedicated to data handling for one of these newspapers – following the model of the Intranet explained above – it is not possible to draw a firm conclusion without IP details.

4.6 STATISTICAL DATA

Literature and reports reviewed are agreeing to put statistical data among the datasets the most re-used, even just considering business re-users[9] and these are quoted among the high-value datasets in the report proposing a reform of the PSI directive[10]. Therefore, this report intends to focus in more details on the logs related to the statistical data hosted on the Open Data platform.

In Luxembourg, STATEC[11] is the governmental agency in charge of collecting and providing the statistical data for the country[12].

---

[8] See e.g. http://datadrivenjournalism.net/news_and_analysis/open_data_journalism

[9] Characterization study of the infomediary sector, 2012:
https://www.ontsi.red.es/ontsi/sites/ontsi/files/121001_red_007_final_report_2012_edition__vf_en_1.pdf

[10] Annex IIa: http://www.europarl.europa.eu/doceo/document/A-8-2018-0438_EN.html?redirect

[11] https://statistiques.public.lu/fr/acteurs/statec/index.html

[12] https://statistiques.public.lu/fr/index.html

The data provided by the STATEC agency have been harvested by the national Open Data platform. As the tool used to monitor the logs of the Open Data portal is basically providing detailed information only for the 500 first rows for each request, people in charge of the platform have launched a query to extract specific information for the statistical datasets. STATEC is providing 160 datasets, around 20% of the datasets of the Open Data platform.

Indicator	Value
Number of datasets provided by the STATEC	160
Share of statistical data pages viewed on data.public.lu	15.3%
Share of statistical data unique visitors on data.public.lu	12.2%

Table 6: STATEC data on data.public.lu

A manual labelling of the data was also led to provide a comparison ground with the general trend. **From the coverage distribution, it is rather similar to the other contents of the platform**, which is consistent for two main reasons. First, the agency is covering a large range of domains to fulfil its missions. The second reason is linked to the nature of data used and released by government, a lot of them being at a statistical abstraction level. Nevertheless, the patterns are different considering these categories depending on the number of pages viewed, as shown in the graph below. As any kind of labelling, the results are in large part reflecting the choices made. For example, the category 'Justice, legal system and public safety' tends to show little interest for statistical data on justice and security, but the data on roads accidents and road safety are labelled under the category 'Transports' and contribute to the relatively high popularity of this category. The amount of data labelled under the category 'Population and society', that could fit a large majority of datasets, given the statistical lens and the duties of the agency, has been limited in order to avoid a flattened view of the other categories, and is still – and unsurprisingly – at the first rank of the viewed data pages. A noteworthy result is **the low popularity of the category "Education, culture and sport**", whereas a lot of datasets on school results are made available Conversely, one can note the relatively **high interest brought to the category 'Agriculture, fisheries, forestry and food', compared with the share of this category considering all the datasets of the platform**. The interest to proceed the same task for the downloaded datasets has been explained for the general downloads, i.e. downloading may be considered as a deeper commitment. For the statistical data, however, the patterns are not affected.

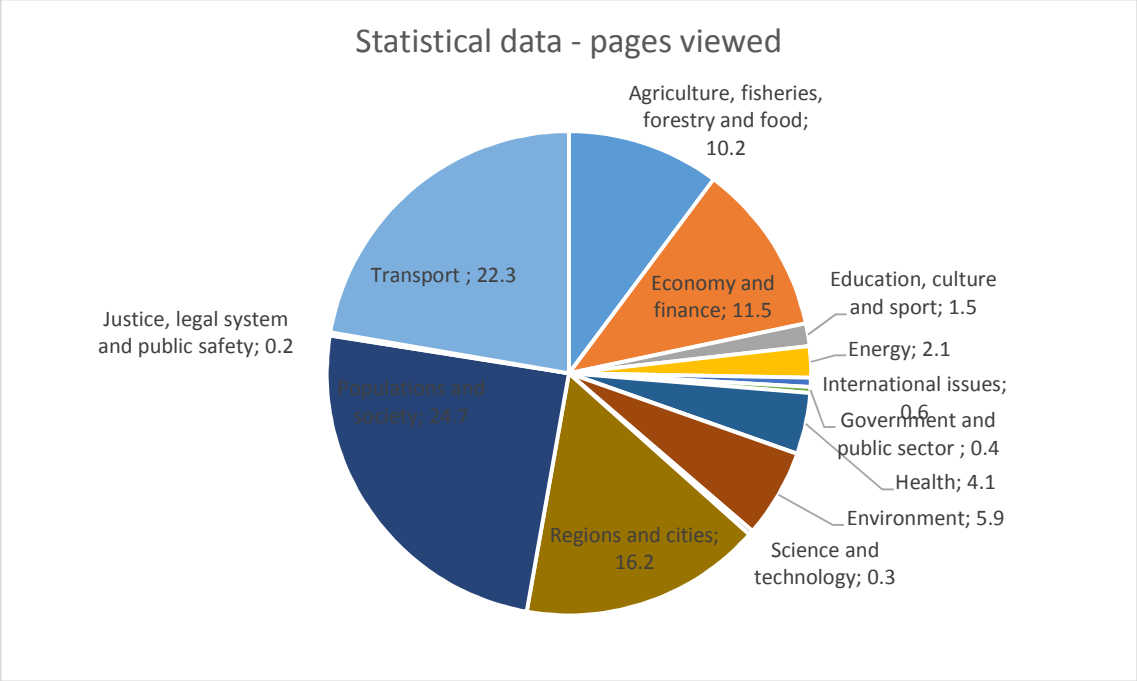


Figure 12 - Thematic distribution weighted by visited webpages

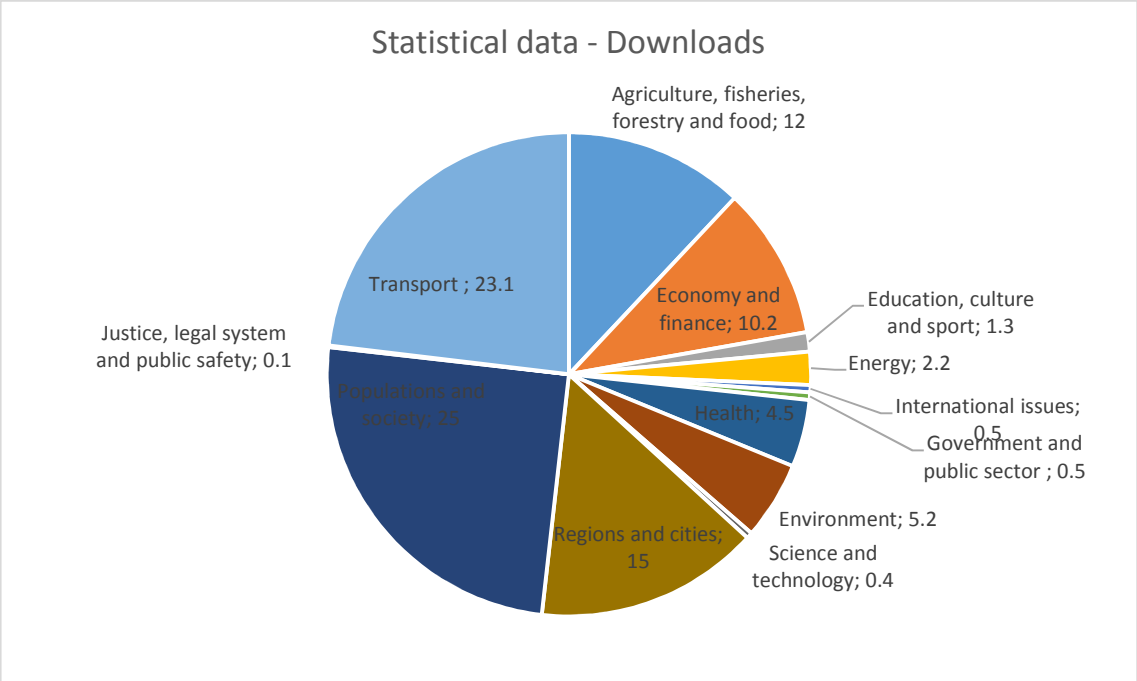


Figure 13 - downloads thematic distribution.

It is not possible to compare exactly the same perimeter between the query logs from the website of the STATEC and the access logs of the national Open Data portal, but it is still providing some

interesting insights, especially as the results are not really consistent: maybe because there are different profiles of re-user depending on the platform.

Concerning the query logs, only search queries (for one month) were available. These data have been labelled following the same EDP typology. Some users, maybe because of the possibility to get relevant answers in generic search engines, are sometimes entering complete questions, e.g. "WIEVIEL % ACKERLAND GIBT ES IN LUXEMBOURG". It does not appear on the data available for data.public.lu, but it's not surprising as the more complex and the longer a query is, the less frequent it will appear, and the dataset is displaying detailed information for the 500 most frequent requests. Compared with the search queries recorded on the Luxembourg Open Data platform, the range is larger. Concerning the searches related to the **economic state of the country, the majority of the queries are related to general indicators** (as expected) such as GDP, salaries, or unemployment at the scale of the country. Nevertheless, visitors are also researching more accurate figures, for example on gender inequalities (e.g. concerning the salaries), search of information about specific companies (especially about their turnover), or time-series of prices for specific products and services. These queries are demonstrating a certain interest for **touristic issues** (3.4% of the queries) that does not appear in the pages viewed on the Open Data portal.

The challenges to interpret query logs in terms of impact has already been explained for the general case. What they are showing here is the **existence of different kinds of impacts, not only from the economic standpoint**. Of course, some of these queries are directly related to topics of interests for a **scientific perspective** in social sciences, such as "pratiques culturelles", i.e. "cultural practices". Unlike the pages viewed on the Open Data portal, where the category related to **justice and security** is more than marginal, it is reaching almost **9% of the queries**, with a large majority focusing on criminality and delinquency. Other queries show also a personal interest expressing a curiosity – but still it is a significant point -, for example seeking information on distribution of babies names.

Indicator	Value
Number of search queries (one moth)	12 046
Number of search queries (annual projection)	144 552

Table 7: search queries: monthly figure and annual projection

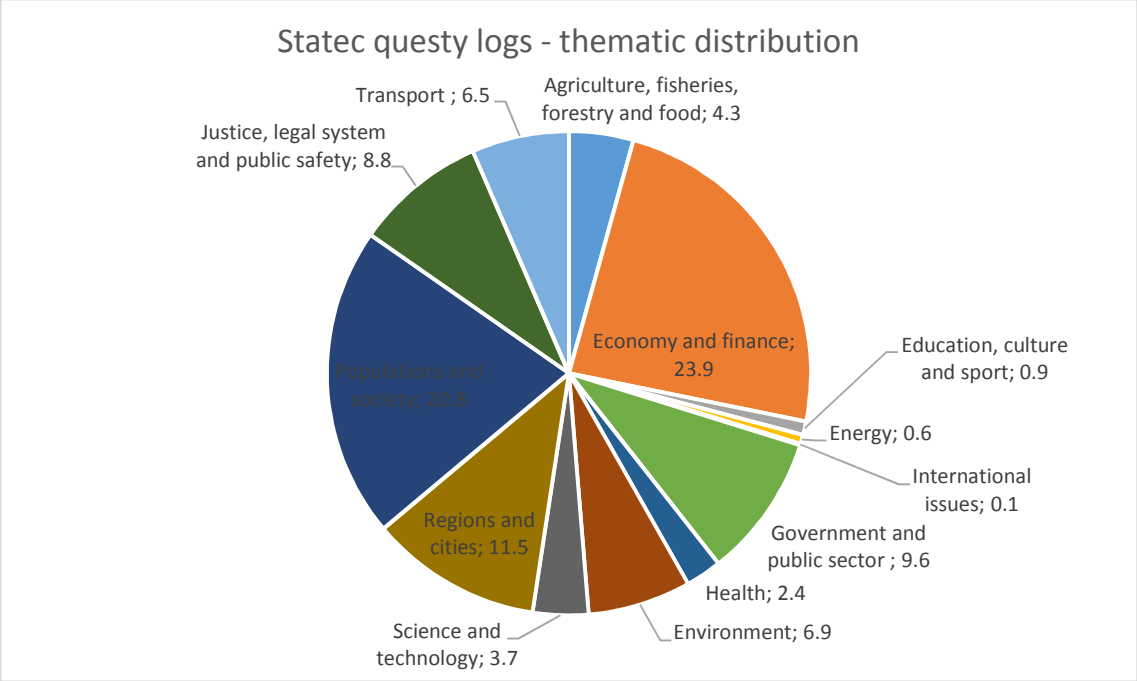


Figure 14 - Search queries: thematic distribution

As for the searches, the majority of the queries are written in French, but the majority is less overwhelming, with 85%. 10% of the requests are expressed in English, the remaining part in German and in Dutch.

If it represents a kind of redundancy, putting the statistical data on the national Open Data portal is contributing to increase the access to these data. One issue to be solved is whether the users are presenting different profiles.

4.7 The Game of Code 2018 Hackathon

Another goal of this study is to assess the contribution potential of logs analysis to gather information on the proper impacts of the actions undertaken to increase the re-use and the impacts of Open Data, such as hackathons. In Luxembourg, one yearly Open Data challenge is sponsored by Digital Lëtzebuerg in Game of Code hackathon since 2016.

Log analysis may bring insights at different moments of the re-use lifecycle leading to the impacts. Upstream (and during the event), one may gather the datasets leveraged by the services built during the hackathon. Downstream, and on the long term, it is useful to consider if it has led to an increase (and potentially re-use) of the datasets, what are the impacts triggered through the release of these services, which is another indirect way of measuring the impacts.

Some of the metrics explored above may be useful to analyze this kind of event. For example, the Website showcasing the hackathon event is at the 21th rank of the known referrers, witnessing the role of the event to bring traffic and potential re-users.

One question to consider is **whether the event organized is affecting in an observable way the platform, reflecting access and re-use**. One possibility is to examine if there are some changes in the patterns of access. Given the data available, in other words, it means checking whether the number of pages viewed during this period is showing significant differences, hypothetically a peak. A specific export added the pages viewed during the four days of the event (15-18th of March, 2018). Monthly and annual exports communicated by the platform and providing consisting figures, allowed to compute an average number of pages viewed per day, and so for a period of four days. This method provides a comparison ground, even rough and fuzzy, showing an **increase of 129.6% of the pages viewed during the event compared with an average period of four days**. The same computation led for the **downloads gives an increase of 109.4%**, the reduced figure might be explained by the fact that re-users are examining more datasets than they are finally trying to input in their product. Of course, a lot of factors may affect the validity of this figure, one which could lead to consider a higher figure being that the challenge is announced in advance and some teams may use this period to prepare their prototype in advance beyond what is computed for these four days. However, the difference with the average, both for downloads and pages viewed, is so large that it remains significant.

On the longer term, different actions may contribute to analyze the impacts. First one is to monitor the created services themselves, in terms of pages viewed, but also on other platforms like *Google Play Store* or *Apple iTunes*. This monitoring has also to take into account the architecture: if the services are embedding the data, if they are updating them on a regular basis, if each (first) use by an end-user is causing an (indirect) interaction with the Open Data platform. Another action is to examine, depending on the topics issued by the organizers of the event challenging the re-users, whether the related datasets are facing an increase compared with the general trend. The table below is summarizing the information brought by the log analysis, upstream or downstream.

Indicator	**Value**
Difference with average pages viewed on the same time windows	+129.6%
Difference with average downloads on the same time windows	+109.4%
Difference with average bounces on the same time windows	+27.5%
Difference with average exits on the same time windows	+30.6%
Number of pages viewed for the services created during the event (temporal coverage: March to December)	751

Table 8 – Impacts of a hackathon on the Open Data platform's attendance

One finding of this report is thus that logs analyses are a suitable way to analyze the actions aiming at fostering the re-use of Open Data, especially for a specific time window, but also in the longer term, in conjunction with other approaches such as interviews.

4.8 SUMMARY

At a statistical scale, access logs of the Open Data portal are showing a large and increasing uptake of the released data. However, some limitations, some contextual (e.g. geographic provenance) some others more structural (e.g. logs data structures) have to be overcome to unleash all the potential benefits of this approach. Analysis of query logs of the platform does not change but confirms the thematic distribution of interest. We showed also that beyond these statistical figures, a careful analysis of logs may provide a basis to lead case studies. 2019 report brings a specific focus on three issues. Referrers allowed to stress the important role of public websites to drain traffic to the platform – identification of the actual users being still a challenge – and to identify the kinds of external sources where the platform could get more influence. Statistical data are released both on the statistical agency and on the Open Data platforms. In spite of this redundancy, a comparison between the access logs (for the Open Data) and the query logs (for the statistical agency) suggest that there are complementarities and that re-users profiles are different. Finally, this report is showcasing the suitability of logs to assess the benefits of a stimulation initiative as the Game of Code hackathon, by considering a time-window around the event, to assess how the platform may reflect (mostly in a quantitative perspective) on short and longer terms the increased interactions implied.

5 OPEN DATA RE-USERS GROUP

On January 25th, 2019, LIST has organized, in partnership with Digital Luxembourg, Interreg NWE project BE-GOOD, Data.public.lu, Luxinnovation and Technoport, the first round of series of workshops dedicated to Open Data in Luxembourg. This first workshop intended to bring together the actors of the mobility ecosystem in Luxembourg and engage exchanges to identify:

- Motivations and obstacles to the reuse of public data
- Needs in terms of public data (availability, documentation, quality, update, etc.)
- Opportunities for cooperation between the public and private sectors in terms of mobility and transport

The workshop gathered a large audience of around 40 attendees coming from the whole Open Data ecosystem in Luxembourg. The executive summary is in annex of this report.

6 CONCLUSION

Initiatives related to digitalization, and more specifically to the Open Data theme, are on the rise in Luxembourg. If this phenomenon is young, and therefore under construction and perfectible, it can also and above all rely on concerned and committed users who, if they describe the weaknesses of the ecosystem, want also and above all firmly to participate in its improvement. The consumer satisfaction survey carried out in this study showed us the willingness of users of the portal to participate in its improvement and the construction of an efficient, effective and sustainable Open Data ecosystem, with precise comments or problems description, always with an idea of solution.

Luxembourg has therefore the huge opportunity to combine Open Data providers and users in the design of the Luxemburgish Open Data ecosystem, in order to achieve the concept of digital nation.

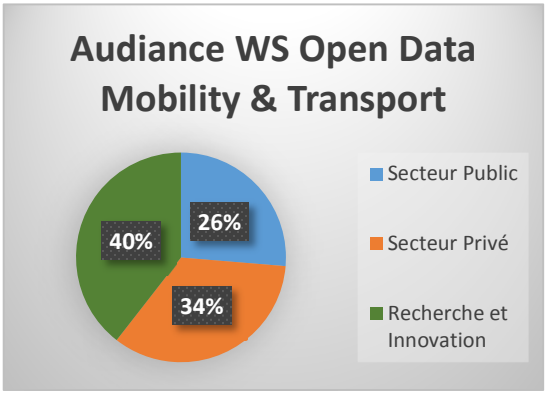
7 ANNEX: "OPEN DATA SEEKING REUSERS"

On January 25th, 2019, the Luxembourg Institute of Science and Technology (LIST) has organized, in partnership with Digital Luxembourg, Interreg NWE project BE-GOOD, Data.public.lu, Luxinnovation and Technoport, the first round of a series of workshops dedicated to Open Data in Luxembourg:

Open Data seeking Reusers

Episode 1 – "On The Road Again"
mobility & transport issues in Luxembourg

In front of a large audience of around 40 attendees coming from the whole Open Data ecosystem in Luxembourg, and after a short welcome speech by Slim Turki, Researcher at LIST, Raymond Feron (Rijkswaterstaat, NL), Program Manager, started the conference with a short presentation of BE-GOOD, an Interreg NEW project, which aims at **unlocking, re-using and extracting value from Public Sector Information (PSI) to develop innovative data-driven services in the area of infrastructure & environment.**



Then, Vanda Turczi (Orléans Métropole, FR) and Edith Nuss (Nexterité, FR) shared their feedback concerning the BE-GOOD challenge "Continuity of Traffic". During the presentation, they highlighted the main steps of their collaboration, from the definition of the scope of the challenge through the involvement of the main stakeholders, the list of their expectations, the identification of data available and data required, to the design of the solution considering the users, their needs and habits, and also the problematic of gathering data in an homogeneous and usable format to fulfil the challenge expectations.

To conclude the conference part of the session, Marina Alleti (Département du Loiret, FR) introduced the challenge "Safer Road", which aims at helping decision makers to objectively steer their actions in the field of road safety, improving coordination and anticipation, and bringing consistency to interactions with the citizens by providing them a better service and better prevention when organizing their everyday life. She explained the different steps to publish a call for tender, and also the difficulties they encountered and how they improved their work for a second publication.

The second part of the workshop was dedicated to discussion about the use and re-use of Open Data for mobility and transport issues in Luxembourg, Including the possibilities of transposition of the BE-GOOD use cases previously presented In Luxembourg.

Considering the **availability – re-usability and needs of open mobility data in Luxembourg**:

We are missing a **centralized incident management tool** (like the one handled by RWS in the Netherlands), that is even more important as the network is partially vulnerable (e.g. few solutions if there is a problem on the motorway to Belgium). There are other problems, e.g. in Kirchberg as there is no traffic buffer.

Police has raw data, transfers them to Administration des Ponts et Chaussées, but for the moment, these raw data may not be published as Open Data, only in statistical form.

Netherlands' representative tells that to be successful, this kind of system requires an easily shareable information, to rally all the relevant providers from the beginning and to encourage their participation highlighting the benefits they could bring from it for their regular duties.

A participant states that roadworks are not coordinated. Administration des Ponts et Chaussées is holding and publishing these data, but only for the part of the network for which they are in charge, 2nd level of roads is not covered. More generally, municipalities do have a high level of autonomy, it would rise some difficulties to introduce that kind of system at the national scale. Even if municipalities data might also raise some interoperability issues pertaining to models or formats, their data are valuable as main and secondary are quickly overloaded.

Moreover, a **crowdsourcing approach** could feed this tool, with social media data, a participation of Police, Post (and other delivery services) to feed a system, as well as more structured data coming from Coyote, Waze or TomTom.

As noted by participants, similar experiments in France allowed to state that around 50% of reported accidents are actually false accidents. Hence the **need of a cross-checking approach**.

Some participants stress also the importance to have **real time data**, for example concerning parking. The latter are already publicly available for Luxembourg Ville. This point shall be stated in the parking management contracts. One shall also take into account whether these facilities do have the right system to produce, handle and share these data. Another participant

stressed the need to have an **alignment of the data format** models, which requires standardization. A good example is DATEX2 which enhances **the re-usability and comparability of data**.

Some participants compare also the data availability with other countries, emphasizing for example that there is a lack of social data (in comparison with the UK), maybe because of the GDPR.

Other kinds of mobility could also be considered. For example, one company is working in Israel to park the scooters. It remains difficult to integrate bicycles & soft mobility in a common system, especially in a data ecosystem.

Replication of the challenge "Continuité du Trafic", for a (test) replication of this BE-GOOD challenge in Luxembourg.

A large range of required data are already available. ACL has also relevant data and is interested by a replication and might participate. Pertaining to the current set of data used, they report some problems with the use of roads-cameras (freeze frames). They gave also an agreement to use (but not to release) Waze and TomTom data. They use to crosscheck their information with Police and Ponts et Chaussées through phone calls.

Use of Open Data in research

National Open Data portal pleads to proceed a fuller linkage of research projects and Open Data they re-use. They emphasize also that as intermediaries, they may help to provide many Open Data in specific topics. Among the mentioned projects, participants mentioned a multi-modal journey planner. Other project aims to integrate all Open Data and put data analytics, especially to build multimodal models to understand how people choose their means of transport, including the use of a tangible table. A national project will envision better plan buses journeys and helping to decide which lines to electrify first. LIST and SnT mention a future project, a decision-making tool.

Acknowledgment

The authors would like to thank all the persons and organizations, from private and public sectors who participated in the various questionnaires or provided data sets.

Authors and contact:

Prune Gautier, Sébastien-Augustin Martin, Slim Turki (*slim.turki@list.lu*),
Luxembourg Institute of Science and Technology (LIST)
IT for Innovative Services (ITIS) Dept.
5, avenue des Hauts-Fourneaux , L-4362 Esch/Alzette, Luxembourg | Tel: (+352) 275 888 – 1 | LIST.lu