December 2021

Authors:

ETC Experts: Yoko Dams, Bart Buelens, Pieter Jan Kerstens (VITO), Holger Berg, Maike Jansen, Jonathan Kirchhoff, János Sebestyen (WI)

EEA Experts: Peder Jensen, Shane Colgan

ETC/WMGE consortium partners: Flemish Institute for Technological Research (VITO), CENIA, Collaborating Centre on Sustainable Consumption and Production (CSCP), Research Institute on Sustainable Economic Growth of National Research Council (IRCrES), The Public Waste Agency of Flanders (OVAM), Sustainability, Environmental Economics and Dynamic Studies (SEEDS), VTT Technical Research Centre of Finland, Banson Communications Ireland (BCI), The Wuppertal Institute for Climate, Environment, Energy (WI), Slovak Environment Agency (SEA)

Further exploration of ‘big data’ and other non-statistical data for CE monitoring

Cover design: EEA

Layout: ETC/WMGE

**Legal notice**

The contents of this publication do not necessarily reflect the official opinions of the European Commission or other institutions of the European Union. Neither the European Environment Agency, the European Topic Centre on Waste and Materials in a Green Economy nor any person or company acting on behalf of the Agency or the Topic Centre is responsible for the use that may be made of the information contained in this report.

**Copyright notice**

© European Topic Centre Waste and Materials in a Green Economy (2021)

Reproduction is authorized provided the source is acknowledged.

More information on the European Union is available on the Internet (http://europa.eu).

European Topic Centre on Waste and Materials   
in a Green Economy

Boeretang 200

BE-2400 Mol

Tel.: +14 33 59 83

Web: wmge.eionet.europa.eu

Email: [etcwmge@vito.be](mailto:etcwmge@vito.be)

Contents

[1. Introduction 1](#_Toc90282592)

[2. Selection process 2](#_Toc90282593)

[2.1. Long list 2](#_Toc90282594)

[2.2. Short list 2](#_Toc90282595)

[3. Prototyping 3](#_Toc90282596)

[3.1. Browser fingerprints 4](#_Toc90282597)

[3.1.1. Introduction 4](#_Toc90282598)

[3.1.2. Data collection 5](#_Toc90282599)

[3.1.3. Data documentation and quality assurance 7](#_Toc90282600)

[3.1.4. Data integration 9](#_Toc90282601)

[3.1.5. Data preparation 9](#_Toc90282602)

[3.1.6. Data analysis 10](#_Toc90282603)

[3.1.7. Summary, obstacles and outlook 11](#_Toc90282604)

[3.2. Car- and bike sharing system – case study Germany 13](#_Toc90282605)

[3.2.1. Introduction 13](#_Toc90282606)

[3.2.2. Data collection 13](#_Toc90282607)

[3.2.3. Data documentation and quality assurance 23](#_Toc90282608)

[3.2.4. Data integration 28](#_Toc90282609)

[3.2.5. Data preparation 28](#_Toc90282610)

[3.2.6. Data analysis 28](#_Toc90282611)

[3.2.7. Summary, obstacles and outlook 35](#_Toc90282612)

[3.3. Web scraping electronic and electrical appliances (EEA) 37](#_Toc90282613)

[3.3.1. Introduction 37](#_Toc90282614)

[3.3.2. Data collection 39](#_Toc90282615)

[3.3.3. Data documentation and quality assurance 43](#_Toc90282616)

[3.3.4. Data integration 47](#_Toc90282617)

[3.3.5. Data preparation 47](#_Toc90282618)

[3.3.6. Data analysis 48](#_Toc90282619)

[3.3.7. Summary, obstacles and outlook 50](#_Toc90282620)

[3.4. Repair data (case studies) 54](#_Toc90282621)

[3.4.1. Introduction 54](#_Toc90282622)

[3.4.2. Data collection 58](#_Toc90282623)

[3.4.3. Data documentation and quality assurance 59](#_Toc90282624)

[3.4.4. Data integration 62](#_Toc90282625)

[3.4.5. Data preparation 62](#_Toc90282626)

[3.4.6. Data analysis 62](#_Toc90282627)

[3.4.7. Summary, obstacles and outlook 68](#_Toc90282628)

[4. Conclusion and outlook 70](#_Toc90282629)

[References 72](#_Toc90282630)

# Introduction

The EEA has been investing, in recent years, in enhancing the knowledge base on resource efficiency and circular economy. The latter is a consolidated and successful labelling for a range of policy initiatives where EEA is putting increasing focus. However, while the concept is consolidated as evident with the recent circular economy action plan, the knowledge around progress and its measurement is still lacking in many respects. While the European Commission’s Monitoring Framework on the Circular economy gives insights from a macro-level perspective based on existing statistical data, many aspects of the circular economy such as remanufacturing, reuse, lifespan of products etc. are currently not covered.

Building on the results of the 2020 FWC on emerging data streams as well as the 2020 work on indicator gap analysis, this activity aims to further explore the use of novel data types including big data, to allow monitoring of aspects of CE that is not captured by the regular statistical dataflows, which tend to capture only the material flow element and miss the more process oriented elements. It will also assist in building the partnerships with specific actors to obtain data regularly.

This research has been investigating which new data streams and underlying methods and procedures are available to show how the current status of the circular economy can be made measurable. Since one of the objectives of the circular economy is to expand the lifespan of products and its parts (Kirchherr et al., 2017, p. 224), retrieving information that could provide an indication of average lifespans of products across Europe is an important aspect. But even more than that, information streams that notify if and in which magnitude the value retention strategies or sharing activities have been taken place in Europe or have been changing our consumption patterns are within the scope of this research.

The results of this work could be used for nourishing new CE indicators and this way serve the circular economy dashboard, which provides information about the current status of the circular economy including the so called ten value retention strategies or R-strategies (EPA Network et al., 2021, p. 9).

# Selection process

As part of 2019 and 2020 ETC work, the ETC/WMGE have been developing a structured overview of blind spots in existing indicator sets to monitor the circular economy. This resulted in the formulation of three priority questions, having the purpose to pool further actions:

* Priority question 1: Are sharing systems of growing importance in business, authorities and consumer expenditures?
* Priority question 2: Are the totality of value retention strategies growing in importance in Europe’s economy (EU/EEA)?
* Priority question 3: Are European consumers switching consumption patterns to circular products or services?

Consequently, a limited number of future indicators to cover some blind spots has been explored and preliminary elaborated. Though data availability seemed to be a major issue. That is where this work offered support.

## Long list

Identifying and selecting the best candidates for prototyping, longlisting and shortlisting work in tandem. As a first step the team identified all possible innovative data streams having the capability to produce information and this way fill the existing data gaps. That is why for each priority question, a list with possible data streams was produced.

## Short list

Out of the long list, a short list of candidates was selected. In preparation of the selection process, all novel data streams have been analysed with respect to a set of criteria being:

* Is the data stream **representing** what it is indicating? (What is the scope: could it cover all EU MS, an entire sector,… or does it represent only a “case”?);
* How easy can we **access** the data? (need to purchase, real-time data or only “once in a lifetime”…);
* How **trustworthy** is the data stream? (Could it be biased because the data stream was aimed to influence decision making (political/strategic);
* Level of **novelty**: using the data stream is considered as pioneering, or only slightly different from mainstream.

Both the ETC teams working on task 4.1.2 and on task 4.1.3, together with EEA task managers Peder Jensen and Shane Colgan, have held a meeting in order to agree on a list of data streams that will be prototyped. For the data streams explored it has been discussed whether a quality check could be done for example by comparing a less trustworthy data stream against one that is more trustworthy or statistical. In case a data stream would produce something highly unsure, it still could offer value for example by deriving trends. In case not the value itself, but the (time) trend is more valuable, an “index” (composite) variable could be created. Information will be created by analyzing trends of (combined) parameters.

Finally, four larger topics were set. For each of them one or more new data streams are considered and the work is fully described in the next section:

* Browser fingerprints;
* Car- and bike sharing system – case study Germany;
* Web scraping electronic and electrical appliances;
* Repair data (case studies).

# Prototyping

Some of the data streams are more experimental, while others are more mature. In order to achieve some sense of consistency despite this diversity, the work descriptions are structured along a data lifecycle model (Blazquez & Domenech, 2018).

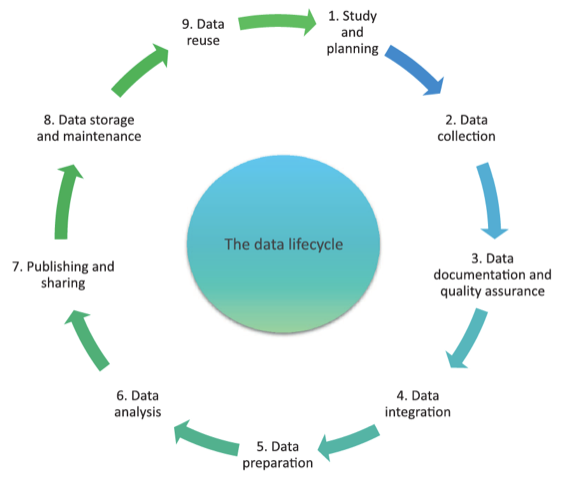


Figure 1 The data lifecycle model (Blazquez, 2018)

In the *data collection phase*, the required data is identified. In the *data documentation and quality assurance phase*, the identified data stream has been described in terms of source of origin, data format, data retrieval process, accessing dates (cf. Blazquez & Domenech, 2018, p. 107), temporal and geographic scope, level of aggregation, features (dimensions / variables) and volume (e.g. number of records). The quality of each data stream was assessed with respect to trustworthiness, currentness, accuracy, completeness, representativeness and interpretability. Once the data is collected and documented, the different data streams were harmonized where needed in the *data integration* phase. In the *data preparation phase*, the data was preprocessed in order to bring it into a format that is suitable for the data analysis techniques applied in the following phase. These steps include the removal of outliers, converting data from string to datetime type, handling missing data, bringing the data into a tabular format and deriving new data from the existing (cf. Blazquez & Domenech, 2018, p. 107), e.g. by means of aggregation, arithmetic or string operations. In the *data analysis phase*, different techniques were applied to obtain and interpret results. A focus was on the description and comparison of patterns extracted from the different data streams. The results were interpreted in relation to the research goal. In the concluding *publishing and sharing phase*, the results of the analysis are reported to the EEA. The generated data stream and source code will be made available to the public in accordance with copyright and privacy concerns.

## Browser fingerprints

### Introduction

#### Objective of this case study and overview

With this case study it has been evaluated if browser fingerprints could be used to determine the age of mobile devices, and hence provide an indication of the lifespan of mobile devices. This in turn contributes to understanding the market situation and the mindset towards the novelty of electronic products.

Before elaborating on the six phases corresponding to the stages of the data life cycle, it is explained what browser fingerprinting is, how it works, by whom it is used and why.

#### What is browser fingerprinting?

“*Device fingerprinting or browser fingerprinting is the systematic collection of information about a remote device, for identification purposes. Client-side scripting languages allow the development of procedures to collect very rich fingerprints: browser and operating system type and version, screen resolution, architecture type, lists of fonts, plugins, microphone, camera, etc.*” (AmIUnique, n.d.)

In general, browser fingerprinting works as follows: When a website is called up, the server stores a wide range of technical features, e.g., which operating system is used, which browser is utilized to call up the website, the size of the device's screen (format) and its resolution (pixels), which colors and fonts are installed in the system, and much more. All these characteristics are combined into a browser fingerprint (Tillmann, 2013, p. 4 and p. 35ff.). Unlike cookies, the browser fingerprint is not stored locally on the user's own computer but at the website provider’s servers (Mann, 2020). The purpose is to allow servers to package data that is appropriate for the particular user environment.

In the following, some examples of information are listed, which are collected in order to determine a browser fingerprint (AmIUnique, n.d.):

* the User agent header
* the Accept header
* the Connection header
* the Encoding header
* the Language header
* the Upgrade Insecure Requests header
* the Referer header
* the Cache-Control header
* the BuildId of the browser
* the list of plugins
* the platform
* the cookies preferences (allowed or not)
* the Do Not Track preferences (yes, no or not communicated)
* the timezone
* the screen resolution and its color depth
* the use of local storage
* the use of session storage
* a picture rendered with the HTML Canvas element
* a picture rendered with WebGL
* Supported Audio formats
* Supported Video formats
* the presence of AdBlock
* the list of fonts

Furthermore, a browser fingerprint is not necessarily unique; for example, two different users may have the same browser fingerprint when using the same computer model and browser, etc. However, the vast majority of browser fingerprints are unique now because the large variety among the devices used, including their various settings and installations results in numerous different browser fingerprint being possible and hence, a very low probability for two devices sharing the same browser fingerprint. Therefore, a browser fingerprint can be assigned to a distinct PC/laptop/tablet/smartphone similar to the IP (Kobusińska et al., 2018, p. 2; Tillmann, 2013, p. 103ff.).1

###### Who uses browser fingerprinting and why?

Browser fingerprinting is used by every major website to ensure good performance, but also to gain insights into user behavior or to establish identity on banking portals or the like. If, for example, the browser fingerprint differs from the one previously detected when logging into a portal, the password must be re-entered and a warning is sent via email (Keenan, 2021). In the case of online stores, information about previous consumer behavior can be used via the browser fingerprint to serve personalized advertising (based on the browser fingerprint, one is identified by the website). In some cases, the data obtained is also passed on to third parties (so-called data brokers), who use it to create a personalized profile and sell it to companies for marketing purposes (Gschwenter, 2021).

### Data collection

#### How to collect relevant data and how to determine the browser fingerprints?

The simplest approach to browser fingerprinting is to construct identifiers by actively and explicitly combining a set of individually non-identifying signals available in the browser environment (Janc & Zalewski, n.d.). Non-identifying signals is information that cannot be used to identify an individual when standing alone, such as the information listed above (e.g., user agent header, accept header). When the information is combined with other non-identifying signals it can lead to a unique combination of information, which couldn then be used to identify an individual. However, these privacy issues are less relevant for us as we only want to know the type of the user devices.

The characteristics of the user regarding their hardware, browser, and/or operating system are read out via JavaScript2.

The data collection itself can be described as stated by Tillmann (2013): When a web page is called up, data relevant to a passive browser fingerprint is stored. When the page is called up for the first time, a unique identifier (uid) can be generated and stored in a cookie. In addition to this unique number, a new database row with a unique row number (id) can also be generated and written to the dynamically generated HTML document. Both data (uid and id) are from now on necessary for the communication between client and server. After the page is fully loaded by the browser and the onload event occurs, active fingerprinting begins. Among other things, a Flash object is used that collects the appropriate data and calls a JavaScript method that sends this data and any other active fingerprint characteristics to the server. Each communication contains the tuple (id, uid). Although the id would theoretically have been sufficient as a primary key to uniquely identify a record, the additional specification of the uid was intended to prevent tampering during transmission. (Tillmann, 2013, p.72f.)3

###### How to get the browser fingerprints from the collected data?

To map data of any size to fixed-size values, a hash function can be used. The values returned by a hash function are called hash values. The simplest approach to combine all attributes into a browser fingerprint is via a hash function pointing to the string where all attributes are concatenated (Kobusińska et al., 2018, p. 2).

Assuming no fingerprint attribute changes on different visits by the user, then such a hash does not differ between successive executions of the algorithm and can consequently be used to identify and recognize a user visiting a web page (Kobusińska et al., 2018, p. 2). At this point, it is important to mention that it is optional and useful for the (most common) purpose of identification. For our purpose, we do not need the hash value, since we are not interested in identification, but in certain individual attributes, and these can no longer be read from a hash value4.

For the collection of browser fingerprints a website is necessary, since browser fingerprints can only be collected if users visit a website. In addition to that, the collection of browser fingerprints requires skills in JavaScript because the characteristics of the browser fingerprints are read out via JavaScript.

Since no website with enough user traffic was at our hands, for the purpose of this study, we chose to use publicly available browser fingerprints instead of collecting them. The Canadian company Inverse inc. developed a platform called Fingerbank, where browser fingerprints of users are collected and even though full access to the data requires payment, it is possible to view 50 browser fingerprints at a time directly on the website (<https://fingerbank.inverse.ca/>).

The data collected from Fingerbank might not represent all mobile devices used, because data is only collected when a user visits [fingerbank.or](http://fingerbank.org/)g, which might lead to biases in the data. However, generating a representative sample size of all currently used mobile devices was not possible anyway due to the limited sample size. Moreover, the purpose of this case study is solely to illustrate the possibilities of using browser fingerprinting data. If this proves to be technically possible, then the next step is to consider how the browser fingerprints can be evaluated as efficiently and scalable as possible.

The data was obtained and transferred manually, so that errors (e.g., extracting the same data twice or missing other data) may occur. Browser fingerprints containing very little information such as no user agent or a very short one was excluded as well as browser fingerprints of laptops which are distinguishable from those of mobile devices. Since all browser fingerprints representing an android device were included, the data represents well the variety of currently used android devices. With regard to the objective of this case study (i.e., to show that it is possible to use browser fingerprints to obtain information about the age of mobile devices), the limited timespan of data collection was not an issue. In fact, it was expedient to not use data collected in different years or months, since we wanted to analyse the age distribution of currently used devices.

#### Using browser fingerprinting to determine the age of mobile devices

Within this case study we implement an exemplary use case to determine the age of mobile devices using browser fingerprinting data provided by the website Fingerbank (<https://fingerbank.inverse.ca/>).

For this case study, our first idea was to combine several pieces of information from a browser fingerprint of a mobile device. However, we quickly recognized that for mobile devices, only the "device name" parameter is needed, which can be extracted from the string attribute "user agent", which is often included in a mobile device’s browser fingerprint. Moreover, it is useful to distinguish between mobile devices and computers or laptops, since for mobile devices only the parameter "device name" is needed for the determination of the device age. However, for computers, laptops and iOS devices this parameter is often not available. So, in this case study, we focus on mobile devices, as age determination is most promising for this kind of device.

The device parameters can usually be linked with a date, e.g., when the model was launched, or when the graphics card was installed for the first time. This means, you can only get an upper bound for the age of the device, i.e., you can only make the statement "the device is not older than x years". Therefore, the age of a device cannot be determined exactly with the browser fingerprint but can only be estimated upwards.

How exactly we use information from the collected browser fingerprinting data in order to determine the approximate age of mobile devices will be described in chapter 2.1.4, 2.1.5 and 2.1.6. Before that, we will first elaborate on the chosen data stream itself in chapter 2.1.3.

### Data documentation and quality assurance

The dataset obtained in the data collection stage and selected for further analysis is characterised in the table below. The description of the data set and the assessment of its quality are based on criteria, which were defined for all case studies within this ETC task. These criteria are divided in two groups: data set description and quality assessment. The data set description includes: source of origin, data retrieval process, storage format, assessing date, geographic scope, temporal scope, level of aggregation, features, volume. The data quality assessment includes: trustworthiness, currentness, accuracy, completeness, representativeness, interpretability.

|  |  |  |  |
| --- | --- | --- | --- |
| **Browser Fingerprints** | | | |
| **Data set description** | **Source of origin:**  <https://fingerbank.inverse.ca/> | **Data retrieval process:**  Manual copy of raw data5  Remark: Data integration is also possible through Fingerbank’s API. However, the access to the entire database of browser fingerprints requires payment. | **Storage format:**  strings / text (manually copied from the website) |
| **Accessing date:**  August 10th, 2021 | **Geographic scope:**  Fingerbank is developed by Inverse (<https://www.inverse.ca/>) which is located in Montreal (Quebec), Canada. No information was found on the geographic scope of the browser fingerprints. | **Temporal scope:**  2014 - 2021  Data collected at all times of the day were used. |
| **Level of aggregation:**  individual fingerprints | **Features:**  user agent incl. device name and version of the operating system | **Volume:**  Number of records: 816 |
| **Quality assessment** | **Trustworthiness:**  Fingerbank is developed by Inverse (<https://www.inverse.ca/>), which is an open source consulting and integration company. No information found on the trustworthiness of this company. | **Currentness:**  possibility to get almost real time browser fingerprints | **Accuracy:**  data itself is accurate  Remark: interpretation of the data only allows for approximate calculation of the age of the mobile devices |
| **Completeness:**  no missing data in our sample of only 81 records  for computer devices and iOS devices attribute “user agent” is missing | **Representativeness:**  The sample of N=81 is rather small and therefore not representative7. Hence, the use case should be repeated with a larger data set, ideally with from major platforms, such as Google for instance. | **Interpretability:**  clear, no ambiguity |

### Data integration

For the purpose of our exemplary use case, data integration was kept as simple as possible. The website <https://www.fingerbank.org/> collects browser fingerprints and provides a limited number free of charge. Fingerbank provides a database of device fingerprints based on various properties that allow applications to fully recognize the device type and then process a custom workflow. Part of a browser fingerprint is the data attribute “user agent”, which is directly visible on <https://fingerbank.inverse.ca/>. Since data access is limited, a big sample size was not achievable, so that it was sufficient for data integration to copy the user agents manually8.

### Data preparation

In the following we elaborate on the data preparation process for our exemplary use case step by step.

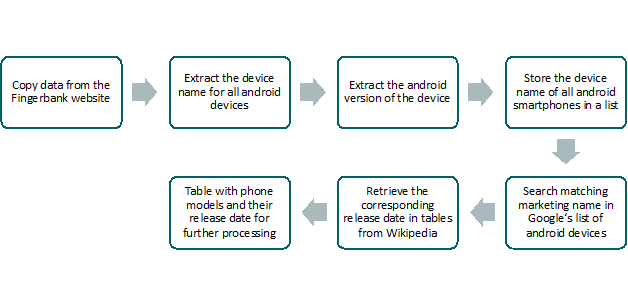


Figure 2 The data preparation process to yield the devices´ release date from its browser fingerprint.

1. The values of the data attribute “user agents” were copied directly from the Fingerbank website (<https://fingerbank.inverse.ca/>) and stored in a list. Raw data consists of a list of strings and needs to be processed (e.g., in Python) in order to gain information.
2. To determine the release date of the mobile devices, the device name was extracted from the user agent string by splitting the string in a suitable way. Not every user agent contained a device name, e.g., apple products only contain the iOS version but not the model.   
   Therefore, *only android devices* were considered.
3. For the 87 android devices in the sample data set, the android version was extracted.
4. In the data sample, there are 5 tablets, 1 android device that cannot be assigned and 4 smartphones that had to be assigned manually. So, in total, 81 Android smartphones were identified. For the 81 android smartphones the device name was extracted.
5. Since the device name is generally not interpretable, the corresponding marketing name of the device was retrieved from a list provided by Google (<https://storage.googleapis.com/play_public/supported_devices.html>).
6. Now that the phone models had been identified, their date of release was determined using tables from Wikipedia that contain the release date for the majority of android smartphones on the market (<https://en.wikipedia.org/wiki/List_of_Android_smartphones>).
7. The release dates were stored in a list and converted from string to datetime type yielding a file listing each smartphone with its corresponding release date.

### Data analysis

Data processing resulted in a list of release dates corresponding to the android smartphones that visited the Fingerbank website (voluntarily and intentionally) with monthly temporal resolution. Since the sample size is small (N=81), a binned histogram is chosen to display the distribution of the release dates (see Figure 3). The bin width is set to three months. In this sample, the average release date is December of 2018 and the empirical standard deviation is 20 months. What can also be seen in Figure 3 is that most of the android smartphones in this data sample are not older than two years.

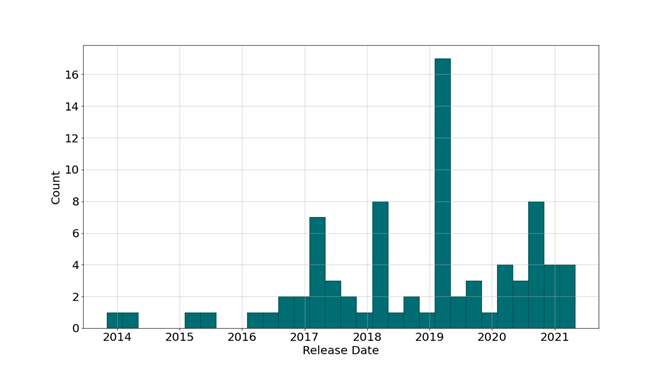


Figure 3 The distribution of release dates of the sample of android smartphones.

In order to analyse to which extent, the software is still up to date in the sampled devices, a histogram of present android versions is created as well (see Figure 4). Even though the majority of the 87 android devices in this data sample run on newer versions of android, 10 devices run on android versions that are no longer supported (5,6 and 7). This means that security patches as well as updates for the operating system are no longer provided and might lead to the conclusion that these devices will soon be replaced and therefore indicates the further lifetime of such a device.

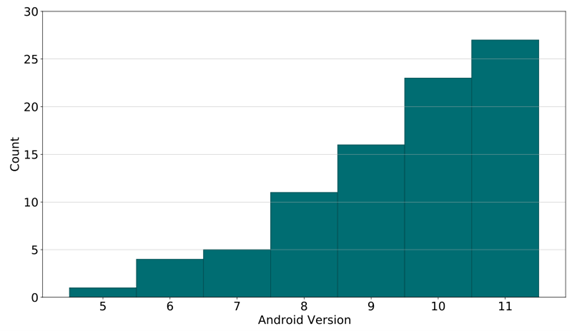


Figure 4 Currently installed versions of android on the android devices sampled.

In terms of what we expected to learn from this analysis, on the one hand regarding emerging data streams and on the other hand with respect to the specific case, we discovered the following. It is technically feasible to use browser fingerprints to determine insights into usage behavior, including the lifespan of mobile devices (i.e., in this case android smartphones). So an initial case for browser fingerprinting as a method to determine the ages structure of devices in use can be made. A time series analysis of fingerprint development can provide insights into the aging and renewal of devices (see also the section).

However, this case study also showed limitations: we do not have a representative data sample in terms of size, and we also cannot draw any conclusions about the users´ location. Thus, these points represent important steps for future research when using browser fingerprints to measure the status of the circular economy and will be elaborated on and summarized in more detail in the following chapter.

### Summary, obstacles and outlook

This case study was the most experimental and exploratory within this ETC task and shows that it is technically feasible to use browser fingerprints to determine the approximate age of mobile devices. Due to limited data access, only a small sample size could be generated. Hence, this exemplary case study is limited to android smartphones and only static data was used to exemplify how browser fingerprinting can be useful to approximate the age distribution of current android smartphones and to check whether their software is up to date or not. With access to large databases (e.g., browser fingerprints collected by a popular website or by a browser provider itself), a more representative sample could be analysed to substantially improve the approximation of the distribution of current devices’ release dates. Furthermore, the methodology could then be extended to more sophisticated analyses.

An important use case of browser fingerprinting is the re-identification of users. With access to such temporally resolved data, usage patterns can be analysed e.g., to determine how frequently devices are used and how the usage intensity changes over time. By conducting longitudinal studies (time series), exact life spans of individual devices can be determined, allowing for a specification of the upper bounds estimated in this study.

When implementing the exemplary use case within this case study, several challenges and practical limitations arose. These are listed and briefly described in the following.

1. The sample data set size is rather small. (N = 81)
2. Only android smartphones have been considered because the “device name” parameter is only available for android smartphones. For Apple products, however, security measures taken by Apple lead to browser fingerprints containing less information. This means that the model of the device, for example, cannot be retrieved. For laptops and computers, browser fingerprints, or the data attributes “user agents” to be more precise, do not contain the parameter “device name”. However, attributes and parameters such as the graphic cards can be used to determine the approximate device age. The distribution of the operating system can (similarly to the versions of android) provide an overview of how up to date computers are.
3. Only the upper bound for the estimation of an android smartphones´ age could be determined.
4. Data quality issues:
   * Representativeness: It would be ideal to have access to a database of browser fingerprints from a website provider or browser provider with representative users.
   * Completeness: For non-android devices, especially Apple products, for example, no device name is available
   * Informative value: The location of the user could not be drawn from the available browser fingerprints in the sample data in this case study. However, it is technically possible to retrieve the location from a browser fingerprint, as the following website shows: <https://www.deviceinfo.me/>. Thus, it would be very important in the future to be able to analyse the location at least at the state or city level. Otherwise, the results are not meaningful in terms of geographic scope and we cannot even say whether the user is in Europe or on another continent, since theoretically anyone can access the website worldwide. In addition, it needs to be investigated in the future whether it is possible to find out the actual location of a user, even if the user uses a TOR browser or a VPN connection, for example. Otherwise, this would bias the statistical results of the data analysis.
5. Data integration challenges: Compatibility of different data sources (e.g., convert table from Google and Wikipedia into csv files)
6. Data privacy challenges: In the context of this case study, it has not yet been investigated whether data privacy issues can arise from the use of browser fingerprints in general. In this study, the data was publicly available and we only used the attribute user agent. In the future, for further use cases where more than one attribute is used, it should be examined whether data privacy issues can arise, since browser fingerprints are usually uniquely identifiable and refer to an individual.
7. Data access: Often website providers who collect browser fingerprints are not interested in sharing them because of competitive advantages and maybe also because of data privacy concerns. Therefore, there is a need for a new data culture towards more data sharing while protecting and ensuring the privacy of the individual (e.g., via data anonymization)

In general, the same procedure of analysing browser fingerprints would work for every device that is connected to the internet and could visit a website such as the Fingerbank website (e.g., tablets, smart watches). Further possible research topics include the question how to use browser fingerprinting to analyse usage patterns in general. For this purpose, the main use case of browser fingerprinting (i.e., reidentification) can be exploited in order to determine how frequently, for how long, and at what times, users visit a website, an online shop or similar. In particular browser fingerprints could be used to analyse how usage frequency decreases with increasing device age, to what extent outdated software is a reason for device replacement, and to approximate the average lifespan of devices. However, for this purpose, data from major platforms would be needed.

Summarized, we claim that the following points will help to improve the usability of analysing browser fingerprints for circular economy assessment:

* Get access to a larger and more representative data stream (i.e., access to major platforms)
* Analyse the location of the users
  + Investigate if it is possible to find out the actual location of a user, even if the user uses a TOR browser or a VPN connection
* Check if in the future, for further use cases where more than one attribute is used, data privacy issues can arise, since browser fingerprints are usually uniquely identifiable and refer to an individual
* Apply and test the proposed procedure on other mobile devices, such as tablets
* Analyse usage patterns, e.g., determine how frequently devices are used and how the usage intensity changes over time
* Examine the total number of different browser fingerprints detected on major websites as a proxy for the total number of devices currently in use and therefore, for the (critical) resources currently in the material cycle. With more comprehensive data on the current landscape of mobile devices and maybe using data on their total weight one could even estimate quantities.

## Car- and bike sharing system – case study Germany

### Introduction

The goal of this case study is to assess the possibilities to inform about the development of carsharing by exploiting three different data streams identified within the task at hand: Branch statistics, web search query data and provider data.

Within the ETC task 4.1.2, a number of data needs were formulated for both, car sharing and the sharing of tools & equipment. Considering the possibilities offered by the data streams mentioned above, this case study focusses on car sharing and the following selection of data needs:

* Number of subscriptions of car sharing companies;
* Number of cities where car sharing platforms are operating;
* Car occupancy rate.

Additionally, the following target dimensions are considered:

* Number of car sharing vehicles available;
* Consumers’ general interest in car sharing.

The case of Germany was selected because different data sources on car sharing were identified for this country: branch statistics, web search query data and provider data. Compiling and comparing insights extracted from these data streams results in a higher validity of the results and makes it possible to assess whether data with a high regional coverage can serve as proxy for other member states. Table 1 maps the above-mentioned target dimensions to the different data streams that are applicable to their analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Target dimension | Data stream | | |
| **Branch statistics** | **Web search query data** | **Provider data** |
| Number of subscriptions | X |  |  |
| Number of cities | X |  | (X) |
| Occupancy rate |  |  | X |
| Number of vehicles | X |  | (X) |
| Consumers’ interest |  | X |  |

Table 1 Mapping of data streams to target dimensions that can be analysed using those data streams. The combinations marked with “(X)” did not produce any useful results.

Although traditional car rental services have similarities with car sharing, they are not considered in this case study.

All data was stored on a private cloud storage, enabling collaboration within the project team while ensuring data protection at the same time.

To achieve the goals of the study, six phases were followed, corresponding to the stages of the data life cycle introduced above.

### Data collection

Three different types of data were collected and stored:

* branch statistics: obtained from Bundesverband CarSharing e.V. (BCS), the branch association of car sharing organisations in Germany. The data was downloaded manually from the BCS website (BCS, n.d.-a) as tables in pdf format and graphs in jpeg format;
* web search query data: obtained from Google Trends, a web site that analyses the relative frequency of search terms and topics on Google Search. For the purpose of an exploratory data analysis, the graphical user interface (GUI) on the website (*Google Trends*, n.d.-a) was used to request and visualise the data for different search terms. Once an initial set of search terms was established, the data was downloaded as CVS files.
* provider data for car sharing: downloaded from the open data portal by Deutsche Bahn AG, a German railway company running a car sharing system called “Flinkster” (*Deutsche Bahn AG,* n.d.). The data sets are available as CSV files. Metadata is provided in pdf format.

#### Branch statistics

The branch association BCS provides chart and tabular data on its website (BCS, n.d.-a, n.d.-b). The data provider was contacted and asked for the data in a machine-readable format. As there was no answer, the data was downloaded from the BCS website in the available formats.

Data on the *market development* of carsharing in Germany from 1999 to 2021 is given as time series plots in jpeg format, see Figure 5 and Figure 6. The numbers of registered users and vehicles are given separately for station-based and station-independent carsharing, and in total on a yearly base. We digitised the chart data semiautomatically using the tool WebPlotDigitizer (Rohatgi, 2020), and manually merged the data into a single CSV file. As the values for the last year are given as numbers, the accuracy of the digitisation could be estimated with a relative error of about 0,5 %.

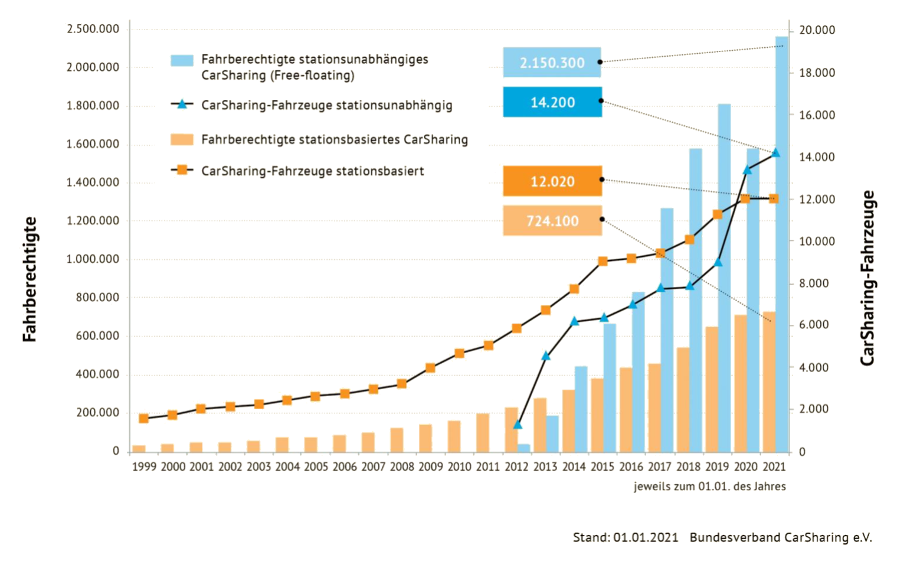


Figure 5 Market development for station-based (orange) and station independent (blue) car sharing in Germany. Registered users (bars, left axis) and vehicles (lines, right axis). Source: (BCS, n.d.-a).

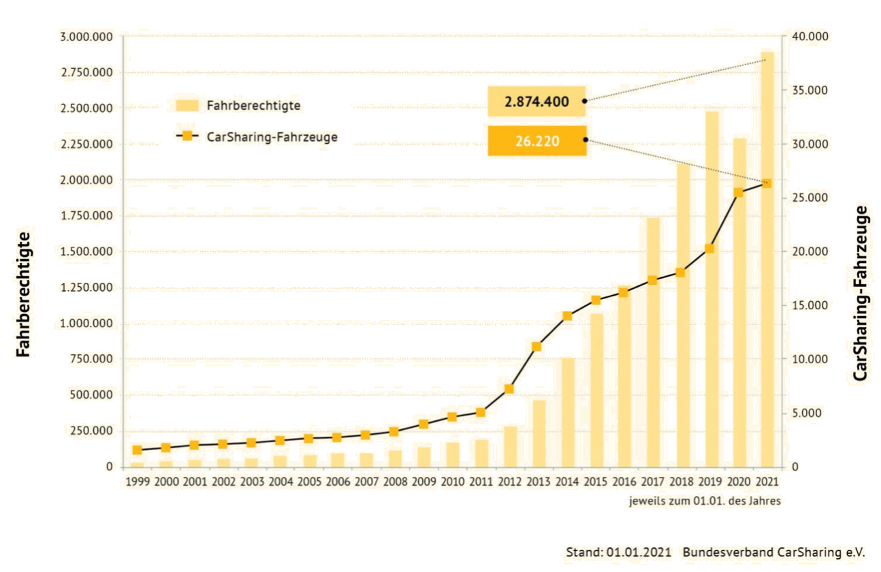


Figure 6 Market development for car sharing (both, station-based and station-independent) in Germany. Registered users (bars, left axis) and vehicles (line, right axis). Source: (BCS, n.d.-a).

*City rankings* are given from 2013 to 2019 bi-annually for cities with 50,000 inhabitants or more (2013: cities with 200,000 inhabitants or more). The ranking is based on the number of car sharing vehicles per 1,000 inhabitants. Also, the absolute number of car sharing vehicles is given per city. Like the market data, the city rankings are given separately for station-based and station-independent car sharing and in total. The data is provided in a tabular format as separate PDF files for each year. See Figure 7 as an example for 2019. The data was manually converted to CSV files and cleansed with spreadsheet operations. Subsequently, the data was merged into a single CSV file with a Python script.

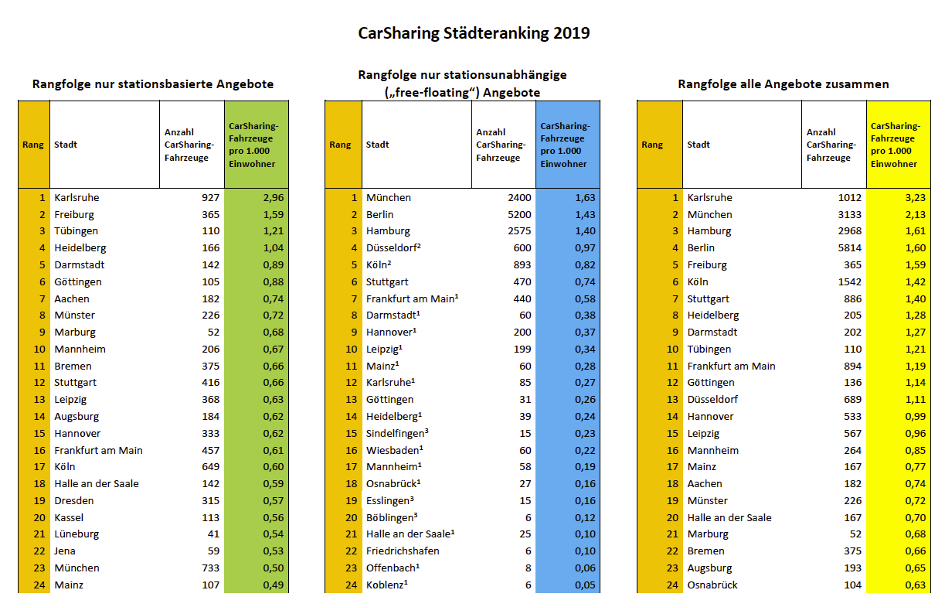


Figure 7 City ranking 2019 for station-based (left), station-independent (center) and total (right) car sharing vehicles in Germany. Absolute numbers (white columns) and relative numbers per 1,000 inhabitants (green, blue and yellow columns), cropped to top 24. Source: (BCS, n.d.-b)

#### Web search query data

In the case of search query data, the first step of data collection is the definition of key words. Following the aim of analysing consumers’ general interest in car sharing, search data relating to either of the following situations are taken into consideration:

* A consumer is looking for general information about car sharing;
* A consumer wants to get acquainted with the local car sharing possibilities;
* A consumer wants to book a car sharing vehicle.

As the same search terms (e.g., ‘Car sharing’) can be used in each situation, it seems impossible to differentiate between the different cases. Also, note that the same search terms might be used in other situations or by other actors (see section “Data documentation and quality assurance”).

Google Trends offers the possibility to analyse search terms or topics. The official help site states: “Search terms show matches for all terms in your query, in the language given”, whereas topics “are a group of terms that share the same concept in any language” (*Google*, n.d.-a). Topics are suggested in the GUI while typing. When analysing search phrases consisting of several search terms, results include searches containing all terms in any order, including combinations with other terms. This default behaviour can be altered by using quotation marks to limit results to the exact phrase, by combining several words with a “+” symbol serving as an OR-operator, or by preceding a word with a “-“ symbol to exclude it from the results. Spelling variants or synonyms of search terms are not automatically included. (*Google*, n.d.-b)

While topics are more general and thus more likely to cover all relevant aspects of a given subject, it is not transparent which search terms are concealed behind a given topic. As transparency is regarded crucial for the given task, we rely primarily on search terms and phrases, and use topics only as a starting point and as a point of comparison to minimise the risk of missing out relevant aspects. Also, search phrases, as explained above, can only be composed by search terms, not by topics.

To find suitable key words, we take an iterative approach, combining data collection and elements of an exploratory data analysis. For this purpose, we use the GUI of the Google Trends website, which allows the graphical comparison of up to five search phrases at a time.

###### First iteration: the topic ‘carsharing’

We start the search for key words with the *topic* ‘Carsharing’ (*Google Trends*, n.d.-b). The geographic scope is set to ‘Germany’, the period of time to ‘2004 – present’, representing the entire available period. Figure 8 shows the resulting time series. Note that the numbers “represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means that there was not enough data for this term” (*Google Trends*, n.d.-b).

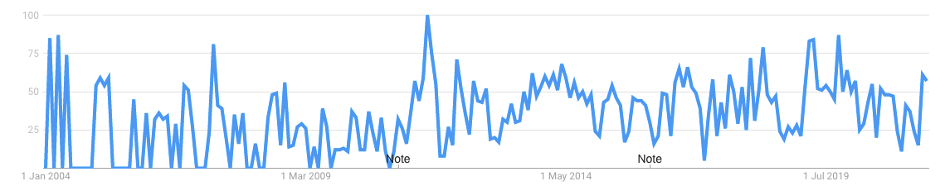


Figure 8 Interest over time for the topic "Carsharing" in web searches using Google Search in Germany. Values are relative to the peak, which is normalised to 100. Source: (Google Trends, n.d.-b).

To get an impression of the topic’s semantics and to get hints for further relevant search terms, we analyse the “related queries” section. This section lists search phrases, which users who searched for the given topic also searched for. Note, however, that these are not necessarily identical with the search phrases comprising the topic. We consider the “Top” metric, which evaluates the most popular related search terms on a relative scale, scoring the most popular search phrase with the value 100. Table 2 shows the related queries for the topic ‘Carsharing’.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Search phrase** | **“Top” metric** |  | **Search phrase (cont.)** | **“Top” metric (cont.)** |
| carsharing | 100 |  | autovermietung | 3 |
| car sharing | 30 |  | mietwagen | 3 |
| berlin carsharing | 11 |  | auto mieten | 3 |
| car2go | 5 |  | carsharing köln | 3 |
| carsharing hamburg | 4 |  | cambio carsharing | 3 |
| sixt | 4 |  | drive now | 3 |
| flinkster | 4 |  | stadtmobil | 3 |
| cambio | 4 |  | car sharing berlin | 3 |
| drive carsharing | 4 |  | carsharing frankfurt | 3 |
| sixt carsharing | 4 |  | share car | 2 |
| flinkster carsharing | 3 |  | carsharing vergleich | 2 |
| drivenow | 3 |  | ford carsharing | 2 |
|  |  |  | bahn carsharing | 2 |

Table 2 Search phrases related to the Topic “Carsharing”. The “top” metric evaluates the most popular related search terms on a relative scale, scoring the most popular search phrase with the value 100. Source: own representation according to (Google Trends, n.d.-b). Colours added by the authors to highlight different content-related groups (see text below).

Looking at the content of the related queries, they can be grouped into five types of search terms:

1. The search term ‘carsharing’ or its variants ‘car sharing’ and – less common – ‘share car’ (blue colour in Table 2);
2. A combination of ‘carsharing’ or ‘car sharing’ and a city name (orange);
3. Brand names or company names of car sharing providers, also in combination with the term ‘carsharing’ (red);
4. Search terms relating to car rental (the German terms ‘autovermietung’ – car rental, ‘mietwagen’ – rental car, ‘auto mieten’ – rent a car) (green);
5. The term ‘carsharing vergleich’ – carsharing comparison (grey).

As a result of this initial analysis, we draw the following conclusions for the further procedure:

* Analyse the *search terms* ‘carsharing’ and ‘car sharing’ and the search phrase ‘carsharing + “car sharing”’ (see below in this subsection).
* Analyse the brand names of the biggest car sharing providers in Germany (see next paragraph).
* Traditional car rental is not considered in this study (see introduction).
* While Google Trends offers the possibility to analyse search interest in different cities or regions, this approach is not taken, as the regional distribution is not a priority for the task at hand.
* The search terms relating to the comparison of carsharing, as well as the rather uncommon search term ‘share car’ are regarded as special cases and are thus not further examined.

A comparison of the search terms ‘carsharing’ and ‘”car sharing”’ and the combined search phrase ‘carsharing + “car sharing”’ with the topic ‘Carsharing’ show the highest accordance for the search phrase – see Figure 9. Therefore, the combined search phrase is used for further analysis.

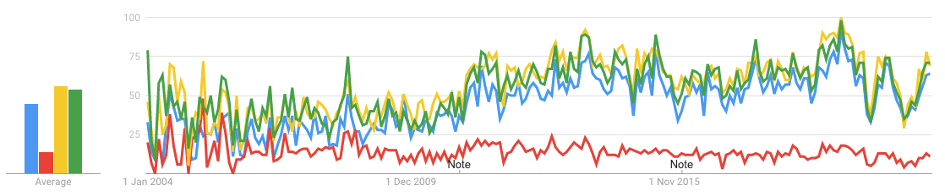


Figure 9 Interest for the search terms ‘carsharing’ (blue), ‘“car sharing”’ (red), “carsharing + “car sharing”’ (yellow), compared to the interest for the topic “Carsharing” (green) on average (left) and over time (right). Source: (Google Trends, n.d.-c).

In a second iteration, search interest for the biggest car sharing providers is analysed. We assume that consumers interested in booking a car sharing vehicle, or get acquainted with a local provider, would often search for a specific provider rather than for car sharing in general.

First, we look at the providers separately to find search terms that optimally reflect the search interest in these providers. Second, we group providers according to whether they offer station-independent, station-based or both types of car sharing. On a methodological level, this approach allows a more detailed comparison with the branch data, which also differentiates between the two types of car sharing. Content-wise, the different types of car sharing may have different implications for sustainability and it is thus interesting to capture how the two submarkets develop.

Table 3 lists the currently biggest car sharing providers by fleet size in Germany, according to (BCS, 2021a).

|  |  |  |
| --- | --- | --- |
| Provider (brand name) | Station-independent | Station-based |
| **Share Now (formerly car2go and DriveNow)[[1]](#footnote-2)** | X |  |
| **Stadtmobil** | X | X |
| **Miles** | X |  |
| **Sixt share[[2]](#footnote-3)** | X |  |
| **cambio** | X | X |
| **WeShare** | X |  |
| **teilAuto** | X | X |
| **book-n-drive** | X | X |
| **Greenwheels (formerly StattAuto CarSharing)[[3]](#footnote-4)** |  | X |
| **Stattauto München** |  | X |
| **Flinkster[[4]](#footnote-5)** |  | X |

Table 3 Biggest car sharing providers by fleet size in Germany. Source: own representation according to (BCS, 2021a).

As in the first iteration, we start with the topic related to the respective provider, where existing, and look for search terms that correlate with it. The following observations are made:

* In the cases of Stadtmobil, teilAuto, Greenwheels and Flinkster (from 2009 onwards), the plain brand name, used as search term, results in a high accordance with the topic for the provider. We use these search terms for further analysis.
* For WeShare no topic is available. Nonetheless, we use “WeShare” as a search term as it appears unambiguous. In the case of Sixt share, there is a topic “Sixt”, but as it may refer to traditional car rental offered by the same company to a large extent, we use the more specific search term “Sixt share”.
* In the cases of book-n-drive, Miles and Stattauto München, topics do exist, but submitting them leads to a technical error. For “book-n-drive” we simply use the brand name and combine it with the spelling variant “book n drive” to a search phrase, assuming that the term is quite unambiguous. The search term “Miles”, however, would also include searches for unrelated topics, such as “km to miles” or “Miles & More”. Therefore, we use the more specific phrase “Miles carsharing”, accepting the risk of missing out some relevant searches which yields an underestimation of the search interest. Similarly, we use the search phrase “Stattauto München”, as “Stattauto” would also include other car sharing providers using the same expression in their names.
* For Cambio, there is a topic, but we could not find a search term or phrase, that yields comparable results: The numbers for the search term “cambio” are ten to hundred times higher than for the topic “Cambio CarSharing”. A look into the related queries section suggests, that this phenomenon relates to the Spanish expression for “exchange rate” - “tipo de cambio”. The interest for the more specific search phrase ‘“cambio carsharing”’, on the other hand, is systematically lower than for the topic, especially from October 2016 onwards (see Figure 10). In the absence of a better alternative, we use this search phrase, knowing that the search interest is underestimated.
* In the case of Share Now, looking at different combinations of “share now”, ‘car2go’ and ‘drivenow’, the highest correlation with the topic “Share Now” is obtained when searching for ‘car2go + “share now”’. This suggests, that one of the merged companies, car2go, is captured by the topic, whereas the other one, DriveNow, is not. This observation underlines, that the untransparent definition of topics make it difficult to interpret the resulting numbers and suggest to rely on well-defined search phrases instead, where possible. In order to analyse DriveNow as well, we use the combination of the three brand names: ‘car2go + “share now” + drivenow’. Figure 11 shows the results for the two alternative search phrases, compared to the topic.

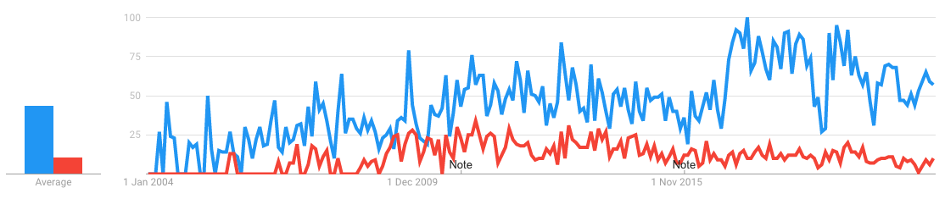


Figure 10 Interest for the topic ‘Cambio CarSharing’ (blue) and the search term ‘cambio carsharing’ (red) on average (left) and over time (right). Source: (Google Trends, 2021a)

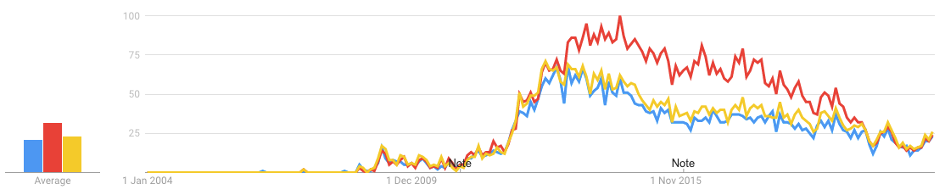


Figure 11 Interest for the topic ‘Share Now’ (yellow) and the search phrases ‘car2go + "share now"’ (blue) and ‘car2go + drivenow + "share now"’ (red) on average (left) and over time (right). Source: (Google Trends, 2021c)

After defining search terms and phrases on provider level, they are combined according to the type of car sharing offered. Table 4 visualises this generation process and its results. The upper part of the table represents the separate queries as described above, whereas the lower part shows the combination for the different types of car sharing (second last row), and the search terms from the first iteration, which serve as a base line (last row).

|  |  |  |
| --- | --- | --- |
| Station-independent | Station-based | Both types |
| **car2go + drivenow + “share now”** | greenwheels | stadtmobil |
| **“miles carsharing”** | “stattauto münchen” | “cambio carsharing” |
| **“sixt share”** | flinkster | teilauto |
| **weshare** |  | book-n-drive + “book n drive” |
| **car2go + drivenow + “share now” + “miles carsharing” + “sixt share” + weshare** | greenwheels + “stattauto münchen” + flinkster | stadtmobil + “cambio carsharing” + teilauto + book-n-drive + “book n drive” |
| **carsharing + “car sharing”** | | |

Table 4 Search terms and search phrases for station-independent and station-based car sharing. Quotation marks are used to combine several words into one expression, while the symbol “+” acts as an OR-operator.

Figure 12 shows the results for the combined search phrases per type of car sharing. The underlying data was downloaded as csv-file for further analysis.

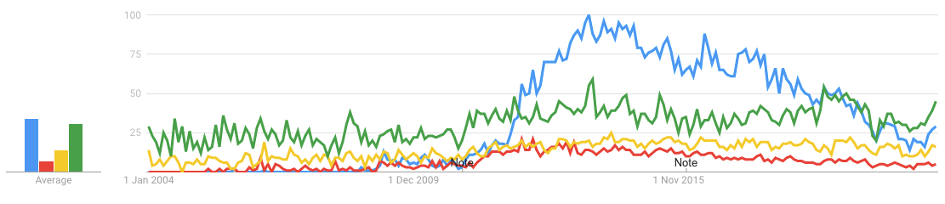


Figure 12 Source: Interest for the brand names of the biggest car sharing providing offering station-independent (blue), station-based (red) and both types of car sharing (yellow), compared to the search phrase ‘car sharing + “car sharing”’ on average (left) and over time (right). Source: (Google Trends, 2021b)

In a third iteration, we looked into spatial data provided in the “Interest by sub-region” section in the Google Trends GUI, where state or city level can be selected, intending to compare the search interest by city with the city rankings of the branch statistics. However, as the output comprised only very few cities we did not pursue this approach any further with regard to the limited informative value.

#### Provider data

Several data sets on the car sharing system „Flinkster“ are offered on the provider‘s open data portal (*Deutsche Bahn AG,* n.d.). The data was published in May 2016 and (partially) updated in June 2016 and in May 2017. Overall time coverage is January 1st 2014 to May 15th 2017, however the data sets on availability and efficiency stop at April 1st 2017 and July 3rd 2016, respectively. Contacted via email, the provider stated, that the data sets on Flinkster are not updated anymore and this is also not planned for the near future. We consider the latest available versions only.

Table 5 lists the data sets offered by the provider and a description of their data content.

|  |  |  |
| --- | --- | --- |
| Data set | Type | Data content (simplified) |
| **Vehicle** | Static | manufacturer, model, registration number, type, engine power, fuel type and capacity for each vehicle |
| **Rental zone** | Static | name, city, country, points of interest, geo coordinates of the car sharing stations |
| **Category** | Static | List of tariff categories, e.g. for different vehicle types |
| **Booking[[5]](#footnote-6)** | Dynamic | booking start and end, driving distance, date of reservation, start and end station |
| **Availability** | Dynamic | availability[[6]](#footnote-7) time (in the sense of duration) per vehicle, day and hour of the day, IDs of the vehicle and the vehicle’s rental zone |
| **Efficiency** | Dynamic | booking time (in the sense of duration) and service booking time per vehicle, day and hour of the day, IDs of the vehicle and the vehicle’s rental zone |

Table 5 Data sets offered by the car sharing provider Flinkster (Deutsche Bahn AG, n.d.).

The data sets on bookings are listed and described on the website, but the download link leads to an error page. The provider was made aware of this and asked if the data could be made available, but did not react. For further analysis we focus on the available data, which is downloaded as csv files.

For the purpose of this case study, the data sets on availability and efficiency, combined with the data set on rental zones, is used to derive the following information:

* Number of available vehicles per year;
* Number of cities in which the car sharing service is offered.

The data set on efficiency, combined with the data set on availability, is used to analyse the following quantity:

* Average occupancy rate (time in use per availability time) of the fleet per year, weekday and time of the day.

The data set on vehicles does not contain temporal information and is thus not suitable to derive the number of cars available for a given period of time. Similarly, the data set “Category” does not contain any information considered useful for the task at hand.

### Data documentation and quality assurance

Each data set obtained in the data collection stage and selected for further analysis is described in a separate profile below. The quality of the data set is assessed based on the quality criteria defined in the study and planning stage.

|  |  |  |  |
| --- | --- | --- | --- |
| **Market development** | | | |
| **Data set description** | **Source of origin:** Bundesverband CarSharing e.V. (BCS), the branch association of car sharing organisations in Germany | **Data retrieval process:**   * Download of raw data (charts in JPEG format) * Digitisation * Merging into a single file | **Storage format:** CSV |
| **Accessing date:** 12th July 2021 | **Geographic scope:** Germany | **Temporal scope:** 1999 – 2021 (station based, total) 2012 – 2021 (station independent) |
| **Level of aggregation:** annual data aggregated over all providers of each type of car sharing (station based, station independent and total) | **Features:**   * number of users * number of vehicles | **Volume:** Number of records (=years): 23 (station based; total) 10 (station independent)  File size: 1 KB |
| **Quality assessment** | **Trustworthiness:** As a lobby organisation for car sharing, BCS might have an interest in reporting high or rising numbers; however the data set is considered trustworthy, as reporting incorrect numbers would risk harming the association and its members. | **Currentness:** reference date (annually): 1st January | **Accuracy**: numerical data apparently rounded to the nearest ten; relative error of the digitisation: 0,1…0,9 % |
| **Completeness:** no missing data (limitations: see ‘Temporal scope’) | **Representativeness:** while BCS represents 185 out of 228 (=81 %) car sharing providers in Germany (BCS, 2021b), the data is requested from all providers (BCS, n.d.-a), thus the data can be regarded as representative | **Interpretability:** clear; note that users of several systems are double counted (see data analysis section) |

|  |  |  |  |
| --- | --- | --- | --- |
| **City rankings** | | | |
| **Data set description** | **Source of origin:** Bundesverband CarSharing e.V. (BCS), the branch association of car sharing organisations in Germany | **Data retrieval process:**   * Download of raw data (tables in pdf format) * Manual conversion into CSV format (one file per year) * Merging into a single CSV file | **Storage format:** CSV |
| **Accessing date:** 14th July 2021 | **Geographic scope:** German cities with 50,000 inhabitants or more (2013: cities with 200,000 inhabitants or more) | **Temporal scope:** 2013 – 2019 |
| **Level of aggregation:** bi-annual data aggregated over all providers of each type of car sharing (station based, station independent and total) for each city | **Features:**   * number of vehicles * number of vehicles per 1,000 inhabitants * rank of the city according to the previously mentioned feature | **Volume:** Number of records (= values for a city): 1014 (total); distribution:   |  |  |  |  | | --- | --- | --- | --- | | **Year** | **Station based** | **Station independent** | **total** | | **2013** | 38 | 9 | 38 | | **2015** | 135 | 18 | 136 | | **2017** | 144 | 22 | 144 | | **2019** | 151 | 28 | 151 |   Total file size: 27 KB |
| **Quality assessment** | **Trustworthiness:** As a lobby organisation for car sharing, BCS might have an interest in reporting high or rising numbers; however the data set is considered trustworthy, as reporting incorrect numbers would risk harming the association and its members. | **Currentness:** published bi-annually | **Accuracy:** unrounded integers |
| **Completeness:** missing data: number of vehicles (total) for 2013 (further limitations: see ‘Temporal scope’ and ‘Geographic scope’) | **Representativeness:** while BCS represents 185 out of 228 (=81 %) car sharing providers in Germany (BCS, 2021b), the data is not limited to those, thus it can be regarded as representative (further limitations: see ‘Geographic scope’) | **Interpretability:** clear, no ambiguity |

|  |  |  |  |
| --- | --- | --- | --- |
| **Search interest** | | | |
| **Data set description** | **Source of origin:** Google Trends | **Data retrieval process:**   * Iterative definition of search phrases * download of CSV-file via web GUI | **Storage format:** CSV |
| **Accessing date:** 26th August 2021 | **Geographic scope:** Germany (applicable to other countries, regions or cities) | **Temporal scope:** 2004-2021 |
| **Level of aggregation:** monthly, country level data | **Features:**  relative search interest for each of the defined search phrases | **Volume:** Number of records (=months): 212 File size: 4 KB |
| **Quality assessment** | **Trustworthiness:** high, as there are no known reasons for incredibility | **Currentness:** data is going back up to 36 hours before the search (Google, n.d.-c) | **Accuracy:** integers on a scale from 0 to 100 |
| **Completeness:** no missing data | **Representativeness:** being by far the most widely used search engine (Desjardins, 2018), Google Search can be considered representative for web searches. By its own account, Google Trends is based on a representative sample of Google searches (Google, n.d.-c). A selection bias towards internet usage has to be taken into account. | **Interpretability:** interpretable as general public interest for a given topic; limitations:   * no possibility to distinguish between different reasons for the search and roles of the searcher (e.g. consumer, investor, citizen) * data may also reflect irregular search activity such as automated searches or searches aiming at manipulating search results (Google, n.d.-c) * non-transparency of the data generation process * normalisation allows for relative comparisons only |

|  |  |  |  |
| --- | --- | --- | --- |
| **Availability, Efficiency, Rental Zones** | | | |
| **Data set description** | **Source of origin:** Deutsche Bahn Connect | **Data retrieval process:**  Download of CSV-files (separate files for availability, efficiency = booking times, and rental zones) | **Storage format:** CSV |
| **Accessing date:** 30th August 2021 | **Geographic scope:** Germany | **Temporal scope:** 1st January 2014 – 1st April 2017 (availability)  2nd January 2014 – 3rd July 2016 (efficiency) |
| **Level of aggregation:** hourly, vehicle level | **Features** (selection)**:**   * availability time per hour of the day * booking time per hour of the day * vehicle ID * city (via rental zone ID) | **Volume:** Number of records (= days x vehicles): 797,655 (availability time) 716,880 (booking time) Total file size: 684.5 MB |
| **Quality assessment** | **Trustworthiness:** high credibility due to fine granularity of the data and transparent documentation | **Currentness:** not updated any more, and no updates planned for the near future | **Accuracy:** seconds |
| **Completeness:** availability: complete, apart from one entirely empty column (in which the id of the key used is stored)  Efficiency: complete, apart from two entirely empty columns (one for the key id and one for the vehicle count, which is described as irrelevant in the documentation)  Rental zones: 24 missing values that were identified as missing geographical identifications (Longitude and Latitude) of car sharing stations, which are not relevant for the case study. | **Representativeness:** of exemplary nature; data is limited to Flinkster, the currently the second biggest car sharing provider by customers with a market share of about 9 % in Germany (Statista, 2020), focussed on station-based car sharing | **Interpretability:** clear, no ambiguity |

### Data integration

As described in section “Study and planning”, an integration of the different data sets on entity-level is not feasible. However, different data sets are compared in the Data analysis section in order to evaluate commonalities and divergences in trends derived from different data streams. To enable these comparisons, the data sets need to be harmonised. Table 6 shows which data sets are used to evaluate which target dimensions, and which harmonisation measures need to be taken to enable comparisons within and across target dimensions.

|  |  |  |  |
| --- | --- | --- | --- |
| Target dimension | Data streams | Data sets | Harmonisation requirements for comparisons |
| **Number of cities** | Branch statistics | City rankings | - Evaluate data for station-based car sharing |
| Provider data | Availability, Efficiency, Rental Zones | - Limit to cities with 50,000 inhabitants or more (2013: cities with 200,000 inhabitants or more)  - Aggregate: yearly  - Note limitation to a single provider |
| **Number of subscriptions** | Branch statistics | Market development | - (none as only one data stream is considered) |
| **Number of vehicles** | Branch statistics | Market development | - Evaluate data for station-based car sharing |
|  | Provider data | Availability | - Aggregate: yearly  - Note limitation to a single provider |
| **Consumers’ interest** | Web search query data | Search interest | - Aggregate: yearly for comparison with branch statistics |
| **Occupancy rate** | Provider data | Availability, Efficiency | - (none as only one data stream is considered) |

Table 6 Data sets used to evaluate the target dimensions and harmonisation requirements for comparisons within and across target dimensions.

### Data preparation

To prepare the data sets for analysis, they were first checked for null values. These were not an issue, since they were either irrelevant or related to technical reasons. E.g., cells were left empty for binary variables instead of being filled with 0 or columns were left empty on purpose when regarding sensitive data such as key ids for vehicles.

Second, irrelevant columns were removed, values were converted into a uniform format, and different data sets were merged. In some cases, new features were created using arithmetic and set operators.

### Data analysis

#### Number of cities

The number of German cities where carsharing is available was obtained from the data set “City rankings”. Figure 13 shows the development from 2013 to 2019 for station based and station independent carsharing. Both quantities show a linear growth, while the slope for station based carsharing is approximately six times as big as for station independent car sharing. The fact that the combined numbers for station based and station independent carsharing is almost congruent with the numbers for station based carsharing alone shows that there are only very few cities where only station independent car sharing is offered.

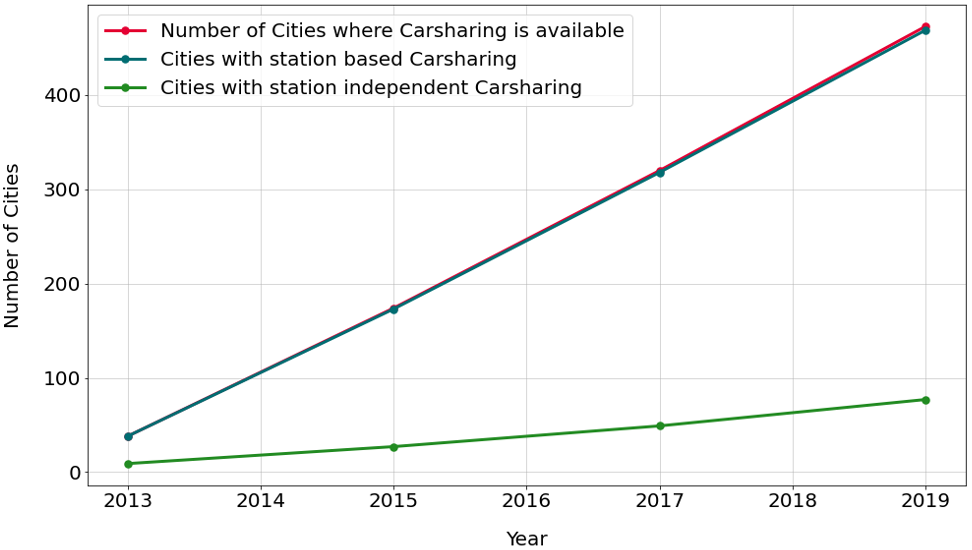


Figure 13 Number of German cities where car sharing is available, derived from the data set "City rankings". In the data set, only cities with more than 50,000 inhabitants (2013: more than 200,000 inhabitants) are considered. Data source: BCS.

As a complementary approach, we analysed the data sets “Availability” and “Efficiency” in combination with “Rental Zones” to determine the number of cities where the provider Flinkster offers its services. However, this approach did not yield any useful results due to irregularities in the data: While the provider seems to have stations in 103 different cities, the number of cities where bookings are made remains at a range of 50 to 70. Also, approximately 10 % of the stations referred to in the “Efficiency” data set are not listed in the “Rental Zones” data set, preventing to match them with a city.

#### Number of subscriptions

The number of subscriptions for carsharing systems is obtained from the data set “Market development”. Figure 14 shows the evolution from 1999 (station independent car sharing: from 2012) to 2021. Both forms of car sharing are growing in numbers. However, while the user base of station seems to approach a level of saturation, the number of users of station independent car sharing outnumbers the first-mentioned shortly after its introduction and shows an upward bending curve. The slight deviation of the station based and total numbers before 2012 can be attributed to a loss of accuracy during the data retrieval process (see section ‘Data documentation and quality assurance’).

The drop of user numbers in 2020 can be explained by the fusion of the two market leaders “Car2go” and “DriveNow”, which merged their customer data, now operating under the brand name “Share Now” (Autohaus, 2020). In other words, to a large extent, the decline in 2020 is an artifact, not caused by people unsubscribing from a car sharing service. From a monitoring perspective, this reveals, that statistical data is not free from irregularities: The user numbers communicated by the branch association are apparently (and understandably) the sum of the numbers reported by the providers. Interpreting these numbers requires to consider that users of several car sharing systems are counted multiple times.

One could assume, that the decline in 2020 is also an effect of the Corona pandemic beginning to affect Germany in spring 2020. However, as the numbers are reported on the 1st of January of each year, the effect of Corona can only be examined by looking at the change from 2020 to 2021. Taking this into account, the opposite is the case: A considerable growth in user numbers can be observed during 2020 for station independent car sharing. This can be explained by passengers switching from public transport to car sharing due to the beginning pandemic (ADAC, 2021).

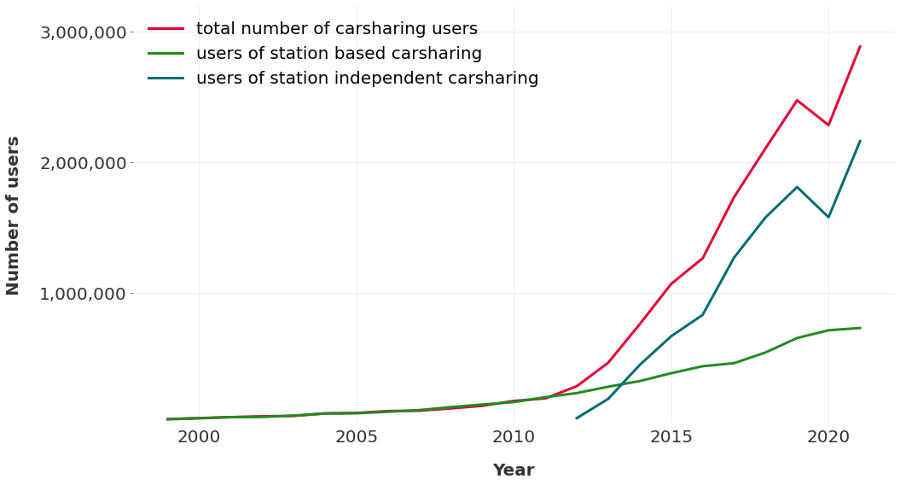


Figure 14 Number of registered users of station based, station independent and both types of carsharing in Germany, obtained from the data set "Market development". Data source: BCS.

#### Number of vehicles

Like the number of users, the number of vehicles was also obtained from the data set “Market development”. The development displayed in Figure 15 shows a pattern similar to the user numbers: For both carsharing types, the number of vehicles is steadily growing over the observation period. Also, the numbers of station based vehicles are levelling off again. Unlike the number of users, however, in the case of station independent car sharing the numbers of vehicles show an s-shaped curve with alternating phases of faster and slower growth. The number of station independent vehicles exceeds the station based vehicles at a later time (2020) than observed for the respective user bases.

The drop of user numbers in 2020 for station independent car sharing contrasts with a pronounced growth of the numbers of vehicles in the same year. This is not surprising, as the merger of the market leaders, that assumedly caused and artificial drop of user numbers has no direct effect on the number of vehicles if we assume that the previously separate vehicle fleets are merged.

Similarly, the growth of vehicles coinciding with the beginning of the Corona pandemic is less pronounced as the incline in user numbers in the same period. An explanation for this could be, that business investments in car sharing vehicle fleets do not respond to changes as quickly as consumer behaviour.

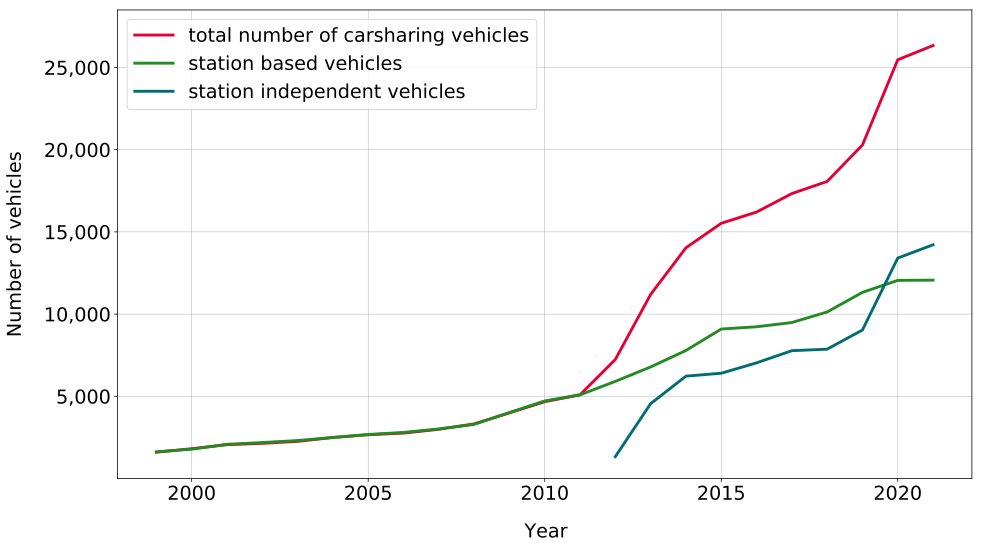


Figure 15 Number of vehicles of station based, station independent and both types of carsharing in Germany, obtained from the data set "Market development". Data source: BCS.

When relating the number of vehicles with the number of users, we observe a growing number of users per vehicle for both carsharing types. On average, station independent car sharing systems show a higher and fast-growing number of users per vehicle than station based carsharing since 2013 – see Figure 16. However, this does not necessarily imply a higher occupancy rate of the station independent vehicles, as the user behaviour (frequency and duration of bookings) might be different.

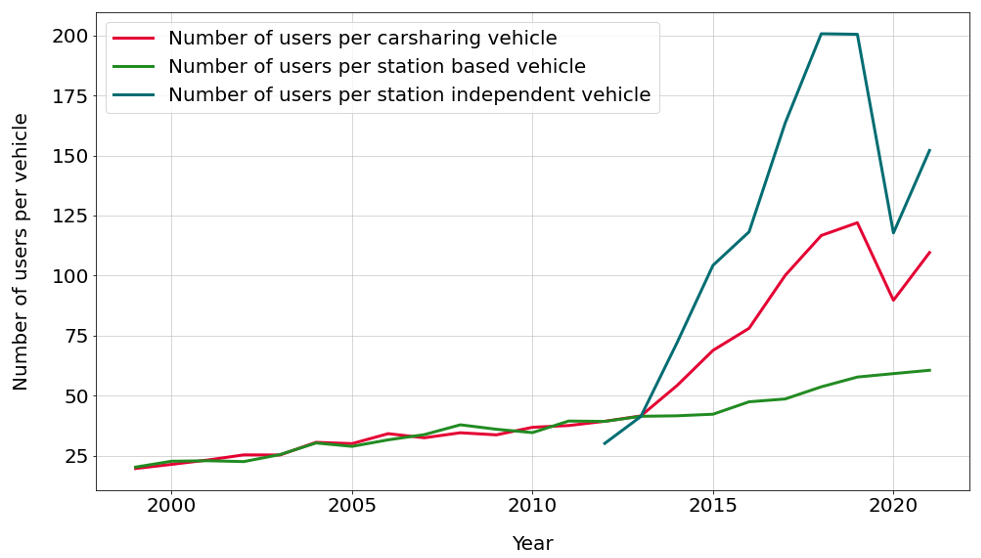


Figure 16 Number of registered users per carsharing vehicle of station based, station independent and both types of carsharing in Germany, obtained from the data set "Market development". Data source: BCS.

As with the numbers of cities, an alternative approach was followed using the “Availability” data set to determine the number of available Flinkster carsharing vehicles. For this the number of unique vehicle IDs was evaluated. Again, data irregularities prevented a meaningful analysis: The number of vehicle IDs fluctuates heavily from day to day, making it unsuitable as an indicator for available vehicles.

#### Consumers’ interest

We analysed the data set “Search interest” to describe patterns of consumers’ search interest for car sharing. Figure 17 displays the search interest for the biggest brands offering different types of carsharing and, as a baseline, the interest for “carsharing” as a key word. A trend curve – the moving average over a 12 month sliding window – is shown in addition to the raw data.

Looking at the trend curves, the search interest for station independent carsharing systems shows the most dynamic development, showing an exponential growth phase peaking in 2014, followed by a gentler decline until 2021. Compared to the founding dates of new providers, no systematic connection can be drawn, as there are examples for both, phases of growth and phases of decline following a company founding. The interest for station based carsharing shows a similar pattern, although less pronounced and peaking already in 2013. The curve for providers offering both types shows the same trend, however in a weak form. The baseline follows a more unsteady pattern with several peaks, showing a slight upwards trend over the observation period.

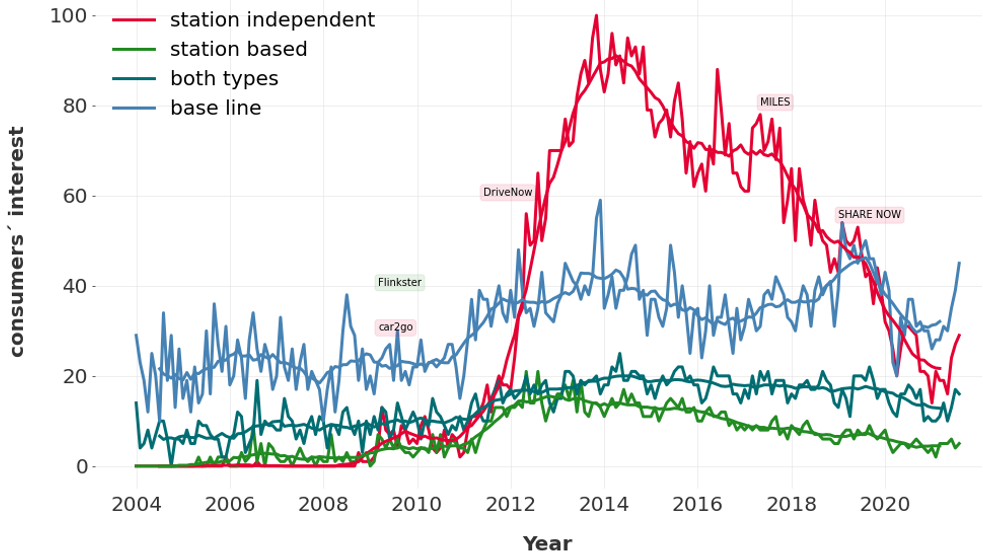


Figure 17 Consumers’ search interest for brand names of carsharing systems offering station independent, station based and both types of carsharing in Germany. As a base line, the search interest for “carsharing” as a search term or topic is given. Data Source: Google Trends. The time of company foundation (found via online research) is labelled for the companies founded within the observation period, coloured by the type of carsharing.

We continued to analyse the data for station independent and station based carsharing in further detail, as they showed the clearest patterns and allow for comparisons with the statistical data. Figure 18 shows the seasonality of the both types of car sharing, averaged over the observation period. One can observe an above-average search interest in March and May to September, peaking in July. This pattern is more distinct for station independent carsharing.

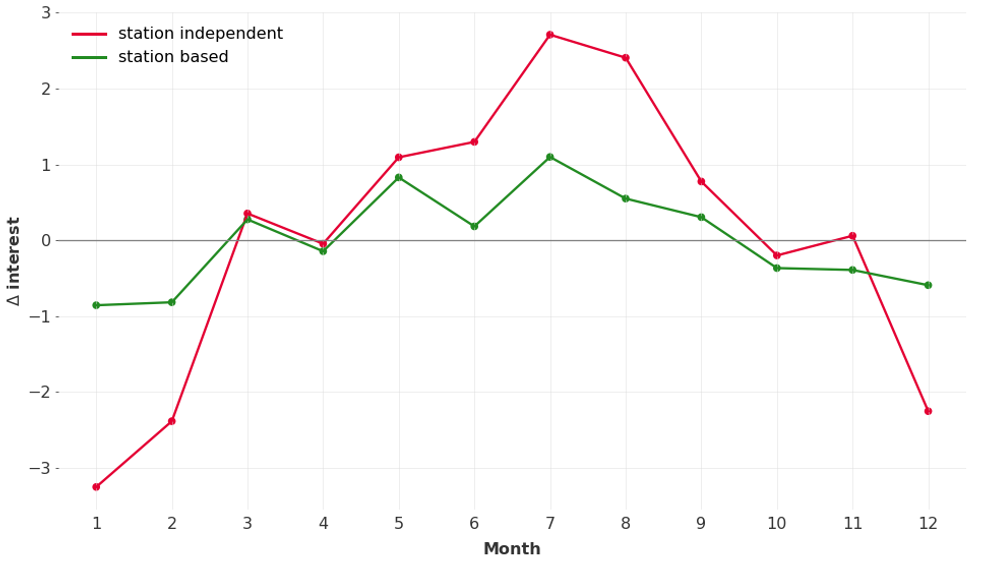


Figure 18 Seasonality of the search interest for station independent and station based carsharing brand names, defined as the deviation of the search interest from the 12 month moving average, averaged over all years of the observation period for each month. Data source: Google Trends.

We finally compared the search interest with the numbers of registered users. While, at a first glance, the steady growth pattern of the user numbers (see section ‘Number of subscriptions’) seem to be unrelated with the peak-shaped pattern of the search interest, we observe a high correspondence when looking at cumulative search interest over time, as shown in Figure 19. In other words: The number of registered users correlates with the overall search interest in the past. This connection becomes understandable when we argue that a certain proportion of users searching for a car sharing provider will register as a new user, while existing users may in most cases access the platform via an app or a website without continuously consulting a web search engine.

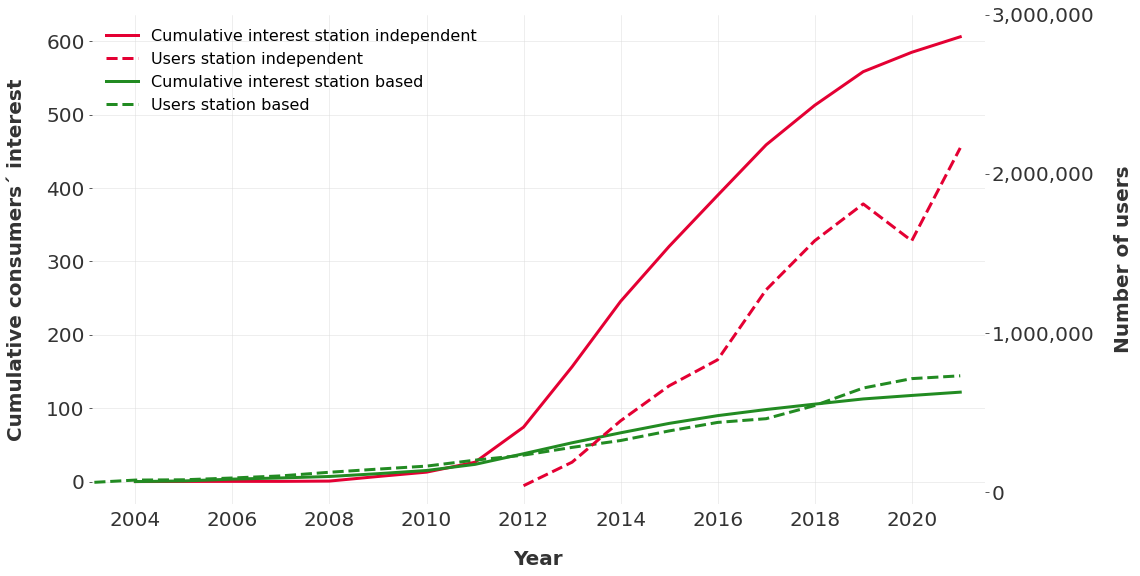


Figure 19 Cumulative consumers’ search interest for station independent and station based carsharing brand names, compared to the number of registered users of station independent and station based carsharing. Data sources: Google Trends, BCS.

#### Occupancy rate

We derived the occupancy rate from the data sets “Availability” and “Efficiency” by evaluating the quotient of booking time and availability.

Figure 20 shows the occupancy rate by weekday and hour of the day. Unsurprisingly, the occupancy rate is relatively low during night-time for all days of the week. Demand heavily increases on weekends and results in occupancy rates of around 40 % in the daytime. During the week, the vehicles are most frequented around noon, especially during the early afternoon, where the occupancy rate lies around 30 %.

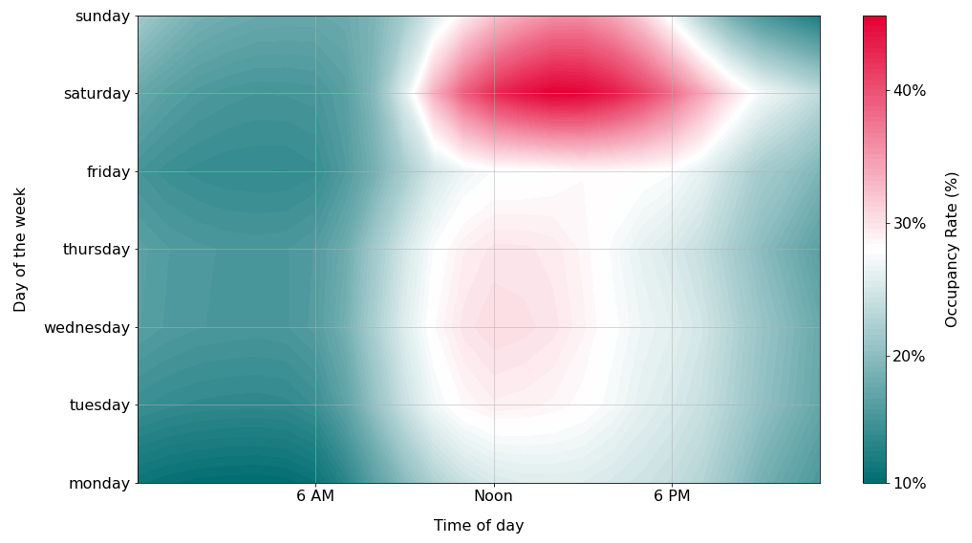
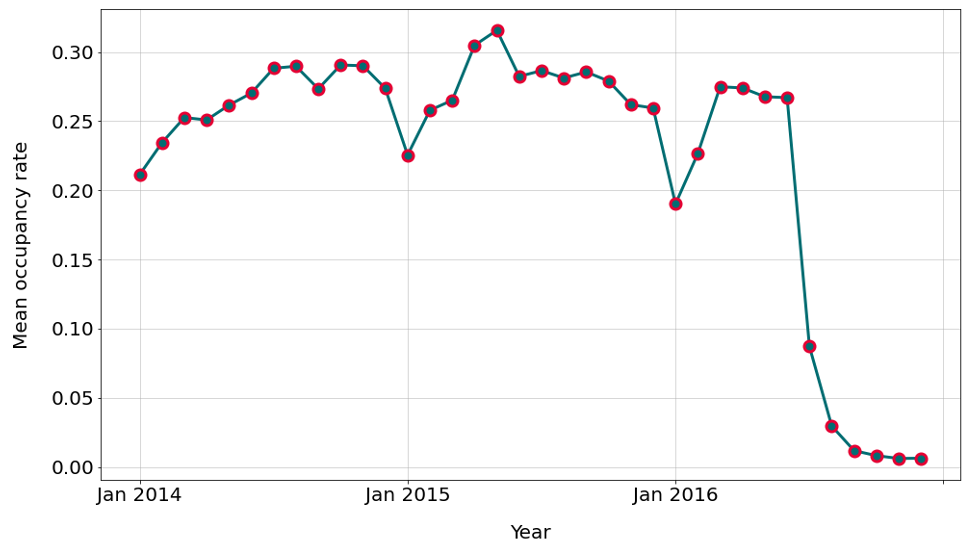


Figure 20 Occupancy rate of carsharing vehicles by Flinkster by weekday and hour of the day, averaged over the observation period. Data source: Deutsche Bahn Connect.

In Figure 21, the occupancy rate is evaluated over the course of the observation period. It fluctuates around the value of approximately 25 % from January 2014 to June 2016. The seasonality pattern is similar to that of search interest (cf. 2.2.7.4): Higher occupancy rates during the summer month and the lowest numbers in January – a pattern that needs further investigation to interpret as one might expect that car sharing use is higher in the winter with weather conditions favouring car sharing over cycling. The extreme drop in the second half of 2016 to near zero cannot be explained by any substantive reason. Looking into the datasets, there are complete records for the whole observation period, with availabilities remaining relatively constant, while the booking times drop in the third quarter of 2016. We assume a change in data collection and/or processing towards the end of the reporting period.

Figure 21 Average occupancy rate of carsharing vehicles by Flinkster over the observation period. Data source: Deutsche Bahn Connect.

### Summary, obstacles and outlook

In this case study, we explored the possibilities to use different data streams to inform about the development of carsharing using the example of Germany. The results can be summarised as follows:

* **Branch statistics**, where available, are a useful and reliable source of information for the monitoring of cities with carsharing, user numbers and vehicle fleets. However, caution is advised when it comes to interpreting the data, as could be shown for the case of user numbers.
* **Web search query data** promise to be a valuable complement to the more traditional statistical data. First findings indicate that search interest can serve as a leading indicator to inform about the development of user numbers. However, further analysis is needed to confirm this finding. From a technical perspective, an optimal indicator needs to be defined. Strengths of this data set are its availability over large spatial and temporal scales, its high resolution and its currentness. However, its interpretability is limited, requiring calibration and validation with other data streams.
* **Provider data** can be useful to detect usage patterns, as shown for the case of occupancy rates. The possibilities this data stream offers depend heavily on the specific data set available and on its quality. As detailed data sets by providers are scarce, the use of this data stream is probably limited to exemplary deep dives.

The **applicability** to other member states differs from data stream to data stream. While web search query data is available for all countries, the availability of the other data sources need to be checked separately for each country. As the web search query data requires to be calibrated with other data sources, its full use can also only be made where some other data is available. The experience of the case study shows that a fair amount of manual work is needed to define key words that match the object of the investigation. Problems occurred with spelling variants, synonyms (in this case rebrandings) and homonyms, some of which could not be solved satisfactorily. As a prerequisite, a list of the biggest carsharing providers operating in the country under consideration is needed.

The approach of analysing web search query data is also applicable to the monitoring of similar topics, such as bike sharing, sharing of tools and equipment, or basically any topic that can be searched for. Complementing statistical data, this data stream can be exploited to obtain more current information (nowcasting) or more accurate predictions of future developments (forecasting). Going beyond monitoring purposes alone, a more accurate forecasting could also be used to optimize fleet allocations in order to achieve a higher occupancy rate and thereby enhance the sustainability of car sharing systems.

**Further research** is needed to develop meaningful indicators from the data streams regarded. A combination of different data streams seems the most promising way to achieve good result by combining the strengths and compensate for the weaknesses each data stream has. The robustness of the findings needs to be validated in different geographical and temporal contexts.

Other approaches to enhance the monitoring of carsharing can also be tested. On a local level, data accessible via APIs for carsharing systems could be used to monitor the number and locations of available vehicles over time, as exemplified in a study of bike sharing systems in Berlin (Agora Verkehrswende, 2019). At a more global scale, download numbers of carsharing apps could potentially be used as a proxy for new users.

## Web scraping electronic and electrical appliances (EEA)

### Introduction

The web is that part of the internet that consists of webpages and websites hosting a wealth of information and data. While overall it is overwhelmingly large and unstructured, in part because of its decentralized nature, the web could contain data or information that is relevant for our purposes and study goals. If and when we find useful data and choose to use it, clearly it was not put online for our purpose, hence we will be repurposing data that was typically put there for other reasons.

Web scraping is a term referring to the automated retrieval of web data, mostly in the form of web pages. Some websites facilitate such retrieval for programmed procedures by providing APIs through which obtaining data that is also shown on websites is made accessible in a convenient, machine readable format, typically JSON.

The web as a whole is intractably large, hence we must bring focus to our efforts and apply targeted web scraping with some purpose in mind.

In this study we focus on electronic and electrical appliances, and investigate repair and refurbish separately. In the remainder of this section, we distinguish between:

* Repair videos;
* E-commerce of refurbished electronic items;
* Repair manuals (IFixit).

#### Repair videos

Many instruction videos about repairing specific products are published on online video sharing platforms such as Youtube, which typically host a wide variety of other videos too. Repair videos range from rather generic instructions about how to replace broken smartphone screens, to very detailed instructions about repairing specific components of particular mobile phone models.

Youtube is a video sharing platform owned and run by Google. People and organisations can publish videos on the platform, making them available to others, typically almost the whole world. Individuals posting videos occasionally can do so, however, many people and organisations publish videos regularly, thereby using channels. A channel is basically an area where videos by the same person – like a vlogger – are posted, or where videos by organsiations or other actors are posted. Channels are often thematic.

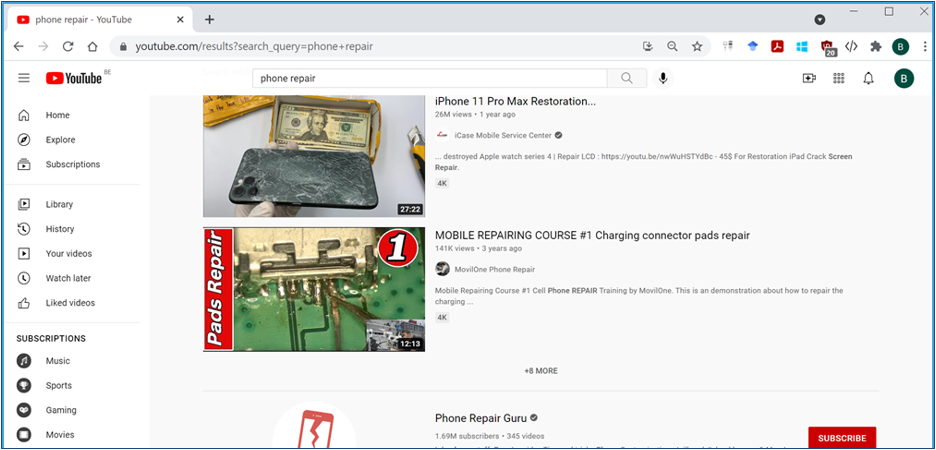


Figure 22 Print screen Youtube webpage

Traditionally Youtube is accessed via their website (www.youtube.com), or via their smartphone app. Youtube viewers can search for videos, subscribe to channels to get notifications of newly posted videos, and many more. Additional functionalities are available to publishers of videos, to provide more detailed insight into their audience, such as how, when and from where their viewers interact with the video material.

Online video platforms expose a search interface through which videos on particular topics can be retrieved, for example on ‘phone repair’. Typically, the number of views of each video is listed, as well as likes, dislikes and comments by viewers.

We conducted an exploration of the possibilities to access and retrieve data from Youtube through its APIs, and investigated the potential use of that data for our purposes. We conducted Youtube searches and a preliminary data analysis to illustrate our findings, and reflect on possibilities and issues.

#### E-commerce of refurbished electronic items

Online shopping is nowadays very popular in the general population and one can buy nearly everything online. While most goods offered online are new, more and more web shops started offering so-called “refurbished” electronic items which have been inspected and repaired before being offered for sale online. These refurbished items may show signs of usage on the outside (e.g. scratches on the smartphone case), but the reseller usually provides some warranty period. Smartphones (and sometimes laptops) are commonly offered online as refurbished items.

The goal of this case study is to assess whether data from online web shops can inform us about the interest in such refurbished items. An indicator that proxies the demand for refurbished items may inform us about the general interest in such refurbished items.

After a first feasibility analysis of different potential online web shops in Europe, we chose to proceed with the Bol.com web shop as a data source. The main reasons for this choice are

* the large market share in Belgium and The Netherlands. Bol.com is one of the biggest B2C online retailers in The Netherlands and Belgium;
* the public availability of a well-documented API to query the product catalogue directly. Other web shops either have no publicly available API (e.g., Coolblue.be) or only provide access to their API in return for referrals (e.g., Amazon.nl).

Bol.com sells a wide variety of products and – for a limited set of electronic products - also refurbished products. We focus on second-hand smartphones and laptops sold by the online retailer bol.com.

Ideally, we would like to directly observe the number of sold refurbished products, but this is not available in the data. The stock of refurbished products is also not available. Therefore, we opted to proxy the demand for a refurbished electronic product by the evolution of its number of reviews.

#### Repair manuals (ifixit.com)

Complementary to the repair videos posted on Youtube, there are websites that offer step-by-step repair guides for electronic devices. This case study looked into the website <https://www.ifixit.com> as a potential data source to gauge the interest in repairing electronic devices.

IFixit offers step-by-step repair guides for a large number of electronic devices: smartphones, laptops, camera's, desktop pc's, cars, tablets and game consoles. These repair guides are community sourced: users can add and edit repair guides themselves.

### Data collection

#### Repair videos

The collection of data pertaining to repair videos is technically achieved through Google APIs for accessing Youtube data. These APIs must be enabled for and from a Google account, via the developer console. One can generate a key that is used to authenticate when querying through the API, a mechanism that allows Google to monitor activity and to sustain a points-based bookkeeping system restricting the number and size of queries that can be run. This is a quotum system where access to the APIs is no longer allowed when the quotum is exceeded. The point-counts are reset to zero every 24 hours.

There are several Youtube APIs:

1. Data API (<https://developers.google.com/youtube/v3>): This API allows searches of Youtube and basic information retrieval. More details below.
2. Analytics and Reporting APIs (<https://developers.google.com/youtube/reporting>). These APIs allow for more detailed information retrieval of videos or channels, however, only for own data sources, which is a serious limitation. This would be needed to get breakdowns of viewers of certain videos by country or by time frame (day, month, ...). In these scenarios, regular reporting would be possible. Since our interest is clearly in videos and channels that we do not own, these APIs are not useful, and we will focus here on the Data API.

Using the Data API, search queries can be conducted programmatically, and the results of the queries are obtained as data sets for further processing and analysis. Search queries can be composed by specifying various parameters, including a query phrase (e.g. ‘phone repair’), the type of materials to search (videos or channels), the maximum size of the results (e.g. 50), the order in which results should be sorted (e.g. by relevance, number of views), a region code to restrict the results to material that is accessible from that region (e.g. ‘BE’ for Belgium), a geographic location and a radius which restricts the result set to videos originating from that circle-shaped geographic region. It is important to note that the views and likes of videos cannot be restricted geographically, and that the geographic origin of videos is not available – hence mapping geographic locations to administrative country or region borders is not possible. In addition it is possible to restrict the search results to videos that were posted in a certain time period.

A query like sketched here returns a list of results, videos or channels, with for example for videos: a unique identifier (that can be used later in other searches), a title, a description, the channel to which it belongs, and publication date. If more details of some or all videos are needed, separate queries need to be run, now retrieving details of specific videos or channels. The returned video-Ids can be used. In this way, additional data can be obtained for videos, including the number of views, likes, dislikes and the number of comments posted about it. Various combinations of searches are tried.

#### E-commerce of refurbished electronic items

The data was directly obtained through the public API of bol.com. One needs to register and request an API key from bol.com to gain access to the API (<https://partnerblog.bol.com/documentatie/open-api/aan-de-slag-2/>). This registration process is fast: we obtained an API key within a day.

The bol.com API provides access to the entire available product catalog of bol.com for both The Netherlands and Belgium. The following actions are possible through different API calls (<https://partnerblog.bol.com/documentatie/open-api/handleiding/api-requests/catalog/>):

* Get product information by searching on specific product ids;
* Get available (temporary) offers for a specific product id from available product sellers. It is possible to filter these offers by “new offers”, “second hand”, “cheapest” or only return the “best offer” available;
* Search product catalog by one or more keywords or EAN/ISBN code. One can limit the search further to specific categories by providing category ids;
* Search for related categories and/or further refine categories starting from one or more specified product categories;
* Recommend products starting from a specific product id;
* Return all related products for a given product as specified by product id. Related products can be accessories or related product families (e.g., the same product in a different color);

The API calls with parameters and returned result format are well documented online. Results of the API calls are returned in JSON or XML format. There is an imposed API limit of 1200 API calls per hour. When this limit is reached the API key is temporarily blocked until a new hour starts and the limit is reset. Furthermore, API calls are throttled to prevent flooding the backend servers: we found that a small pause of 0.25 seconds between successive API calls avoids most of the throttling.

Figure 23 shows what product related information is returned by the API.

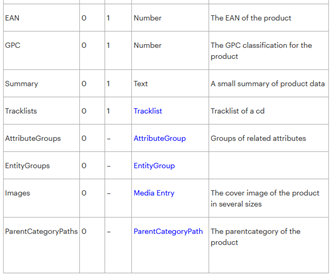


Figure 23 Print screen Bol.com API

Here one sees that both the productid (“Id”) and the average product rating (“Rating”) are available. A product page of, for example, a refurbished Apple iPhone X further provides the number of reviews (highlighted in Figure 24):

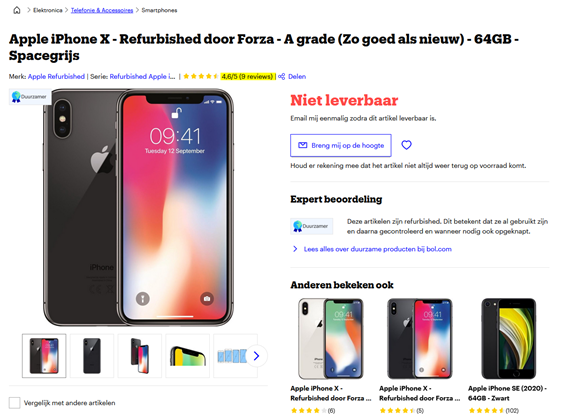


Figure 24 Print screen Bol.com product search

We used Python to extract the information by scraping the product page and searching for the words matching the regular expression “[0-9]+ review[s]?”. Thus, we combined both information obtained from the API with web scraping of the product page to collect product id, number of reviews and average product rating.

We have been collecting this data approximately once a week for 567 refurbished smartphones (category id = 49959) and 12 refurbished laptops (category id = 52680). This represents the full catalog of refurbished smartphones and laptops from bol.com in Belgium at the time of writing.

#### Repair manuals (ifixit.com)

We used web scraping to extract the necessary data for the IFixit website and therefore we did not need an API key.

The IFixit website is organized by starting from 8 large categories: Mac, phone, laptop, camera, desktop pc, car and truck, tablet and game console. Each category is further divided into subcategories. For example: the “phone” category is subdivided into 7 categories (Apple iPhone, Android Phone, Windows Phone, Blackberry Phone, Other OS (Feature) Phone, Phone Accessory and Landline Phone) and the “car and truck” category is subdivided into 52 categories which represent car brands. Figure 25 illustrates this for the “phone” category.

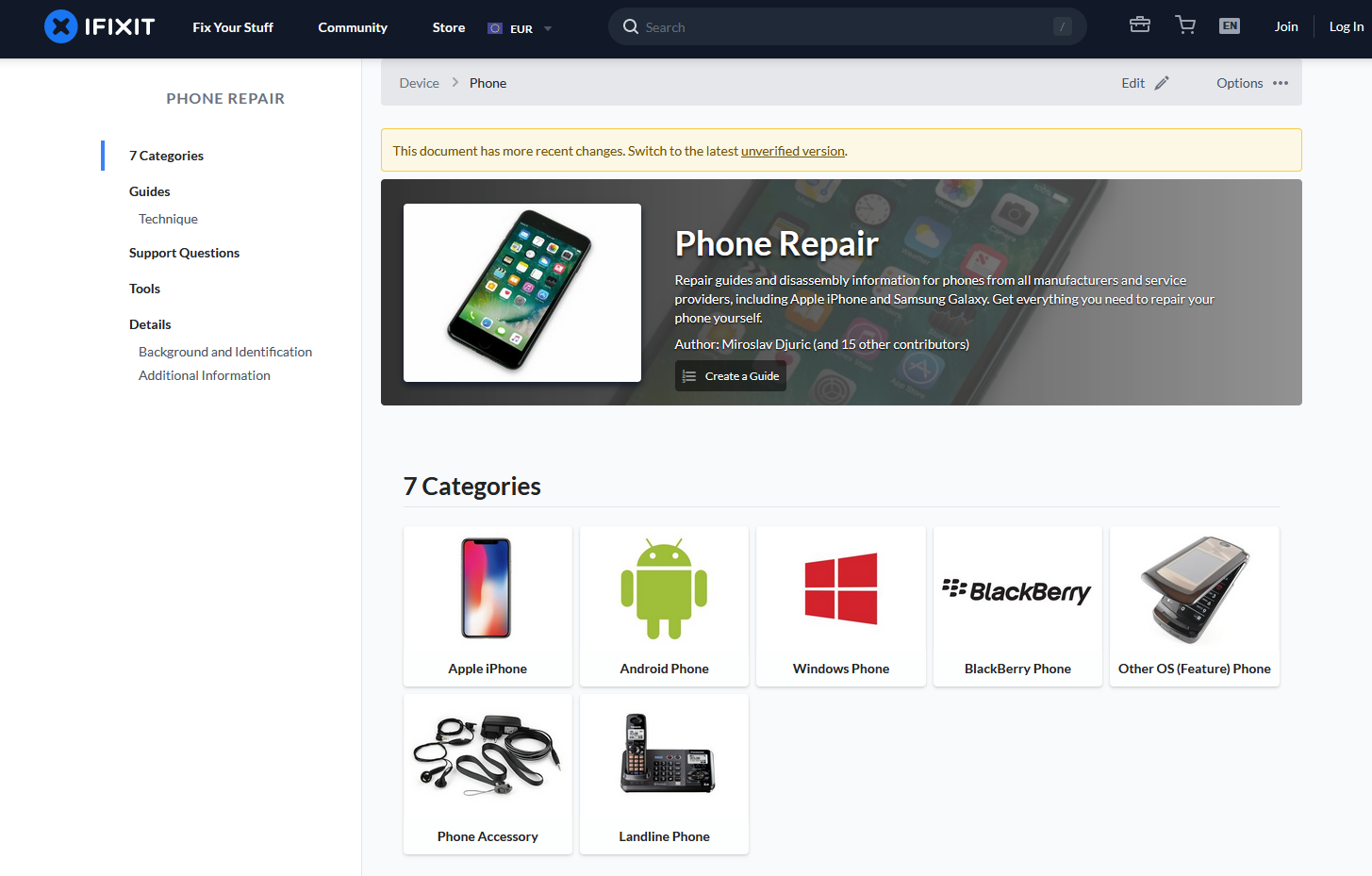


Figure 25**.** The “phone” category and its subcategories on ifixit.com.

These subcategories are then further subdivided into other subcategories and so on. Eventually one arrives at a specific model page (e.g., iPhone 5).

At the bottom of each page there are view statistics for the past 24 hours, past 7 days, past 30 days and all time:



Figure 26 Print screen IFixit

If we assume that most site visitors start on the main website before gradually clicking through the different categories and subcategories to arrive at their specific electronic product, then the view statistics of the 8 top-level categories (i.e., Mac, phone, laptop, camera, desktop pc, car and truck, tablet and game console) capture most page views of the site visitors. For example: a visitor that starts on the main page and ends up at the “iPhone 5” page shows up in both the view statistics of the “phone” category page and the “iPhone 5” page. Thus, it is sufficient to collect the view statistics of the top-level categories.

We use the “BeautifulSoup”-package in Python to scrape the view statistics from these webpages by first searching for the HTML tag of the “stats-numbers-container" class. Within this tag one can then read out the 4 desired view statistics. We have been collecting this data approximately once a week. However, it is equally possible to collect this data at a different frequency.

### Data documentation and quality assurance

|  |  |  |  |
| --- | --- | --- | --- |
| **Repair videos (Youtube)** | | | |
| **Data set description** | **Source of origin:** youtube.com | **Data retrieval process:**  Using python code, through a Google API | **Storage format:** Any, CSV chosen for simplicity |
| **Accessing date:** Multiple days in Oct-Nov 2021 | **Geographic scope:** Global; however, only videos accessible for viewing in Belgium are considered | **Temporal scope:** Videos in test data from 1/1/2017-30/6/2021; earlier data is available too. |
| **Level of aggregation:** Data for individual videos, view counts are aggregated for the video’s lifetime. Number of videos are aggregated by topic but can be disaggregated in time to daily level | **Features:**  1/ search per topic: videos with title, description, publication date, number of views, likes, dislikes and comments.  2/ searching the past based on topic: the number of videos posted in a certain time period. | **Volume:**  The actual volume of data on Youtube is very large, however the number of searches and search results is capped by means of a point-based bookkeeping system controlled through the API. There is a maximum allowance and different actions count for different amounts of usage points. |
| **Quality assessment** | **Trustworthiness:** The actual data retrieved is likely to be trustworthy, in the sense that views and counts are probably correct, although verifying this is impossible. | **Currentness:** Data from Youtube can be very current, as the platform is in constant use and data can be retrieved at any time. | **Accuracy:** Count and view data is probably accurate, but when considering this as a proxy for something else – like interest, or repair actions – it all depends on the correlation between the Youtube data and the quantity of interest. |
| **Completeness:** It is not possible, nor feasible, to retrieve all Youtube data in its entirety. Nevertheless, the data that is retrievable appears to be complete in the sense that it rarely contains missing values. However, some results appeared to be either erroneous or spurious values. | **Representativeness:** Not representative for repair activities in general or repair activities by consumers. There may be other video platforms or places to share repair videos too, in which case Youtube is not even representative for repair video viewing. Nevertheless, Youtube is a major platform, so it probably reflects at least to some extent repair video viewing behavior. | **Interpretability:** Data obtained from Youtube is generally easy to interpret. However, some underlying algorithms remain unknown, for example the basis on which it is decided that a certain video is returned as a result of a particular query. |

|  |  |  |  |
| --- | --- | --- | --- |
| **E-commerce of refurbished electronic items (Bol.com)** | | | |
| **Data set description** | **Source of origin:**  <https://api.bol.com/> | **Data retrieval process:**  Automatic using Python script | **Storage format:**  CSV files |
| **Accessing date:**  Multiple days in October-November 2021 | **Geographic scope:**  The Netherlands and Belgium | **Temporal scope:**  2021 - …  Data collected at all times of the day were used. |
| **Level of aggregation:**  Individual products | **Features:**  Product id, product title, number of reviews, average product rating, product url | **Volume:**   567 refurbished smartphones + 12 refurbished laptops collected at 5 different time periods approximately once per week. The final collected data is small in storage size. |
| **Quality assessment** | **Trustworthiness:**  Trustworthy, but we are unable to check the completeness of the number of reviews number (i.e., we cannot verify that bol.com publishes all product reviews). | **Currentness:**  Data is current: the API returns the info of the current product catalog. | **Accuracy**:  Accurate |
| **Completeness:**  Data represents all available refurbished electronic devices offered by bol.com at time of writing. Missing values are only encountered for the number of reviews, but this simply means that no reviews are available. | **Representativeness:**  Not representative for reuse activities in general as there are other platforms that sell refurbished electronic items. | **Interpretability:**  Data is easy to interpret. The API returns the requested data matching the specified category id. The category name matching the category id is generally self-explanatory. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Repair manuals (ifixit.com)** | | | |
| **Data set description** | **Source of origin:** [**https://www.ifixit.com**](https://www.ifixit.com) | **Data retrieval process:**  Automatic using Python script | **Storage format:** CSV files |
| **Accessing date:** Multiple days in November 2021 | **Geographic scope:** Global | **Temporal scope:** Unknown, but likely very long as, for example, Apple iPhone 1 from 2007 also has a product page. IFixit exists since 2003.**[[7]](#footnote-8)** |
| **Level of aggregation:** View statistics are always aggregated to day, week, month or all time level. View statistics are available for individual electronic devices and encompassing categories. | **Features:**  Product category, view statistics for past 24 hours, past 7 days, past 30 days and all time. | **Volume:**  Small volume of data as only the view statistics for the 8 top-level categories is collected. |
| **Quality assessment** | **Trustworthiness:** Trustworthy, but biased towards more internet-savvy parts of the population. | **Currentness:** Data can be very current, as the platform is in constant use and data can be retrieved at any time. | **Accuracy**: The data itself is expected to be accurate as these are view statistics. One does have to trust IFixit not to fabricate (or adjust) these numbers. |
| **Completeness:** We encountered no missing values for the top-level category pages. | **Representativeness:** Not representative for repair activities in general or repair activities by consumers.  Repair guides are community sourced so the available guides are likely to be biased towards the most popular devices sold on the market and devices of lesser-known brands are probably less likely to have a repair guide available. | **Interpretability:** Data is easy to interpret. |

### Data integration

If one of the data streams should be integrated with other data sources, it will most likely be a matter of matching by date or by time period. Depending on the frequency of the different data sources, some interpolation or aggregation may be necessary.

#### Repair videos

For technical reasons, multiple queries need to be run through the API to obtain a data set relevant for analysis. These reasons include limits on the number and types of queries that are allowed through the API, and the fact that one type of query results in videos, while other types can retrieve data of given videos. Multiple files resulting from such queries can be combined into a single data file or data base for further analysis.

Since the construction of a series covering a longer period (multiple years) requires the combination of search results over that period, we choose to concentrate here on counts of newly posted materials. We therefore setup queries to retrieve data at a weekly basis for the weeks from the beginning of 2017 until mid-2021. Each week results in its own data file. These were summarized and combined into a single file to produce the graph shown in the results below.

#### E-commerce of refurbished electronic items & Repair manuals (ifixit.com)

The collected data at each point in time is stored in a separate CSV file to keep the setup as simple as possible. No further data integration is necessary.

### Data preparation

#### Repair videos

Most of the data preparation needed is the merging of resulting data files. In addition, it is worth checking the data types and contents of particular data fields. In the example data file, the total number of weekly counts appeared unusually high and often but not always the same, unlikely high number, for some time periods. It was decided to remove these values and to replace them by values obtained through linear interpolation from the adjacent time periods.

#### E-commerce of refurbished electronic items

Each CSV file contains the data at a different point in time. We use Python and the pandas library for data analysis and visualization. The data preparation then involves 3 steps:

1. Import all CSV files and merge them into a single dataframe (i.e., a large matrix);
2. Pivot the dataframe in a “wide” format with one column for every product id;
3. Forward fill any missing value along the rows for each column: a missing value is replaced with the last non-missing value in a previous row. The number of reviews cannot decline in normal circumstances so forward filling is a valid approach here for handling missing data. If no review is available for a product id then the missing value is retained for all those product observations.

#### Repair manuals (ifixit.com)

Each CSV file contains the data at a different point in time. We use Python and the pandas library for data analysis and visualization. The data preparation then involves 2 steps:

1. Import all CSV files and merge them into a single dataframe (i.e., a large matrix);
2. Pivot the dataframe in a “wide” format with one column for every product category;

We did not encounter any missing values in the view statistics.

### Data analysis

#### Repair videos

Since a study of temporal changes or trends of views would require building a suitable data set in an incremental fashion, the data analysis in the present context was restricted to a study of the number of newly posted videos. The search phrase ‘phone repair’ was used in the queries, which were set to consider weekly data. The results are plotted in Figure 27. As explained in section 3.3.5, unlikely values were edited.

A long term trend is visible with two periods of relatively higher numbers, and a clear increase at the end, roughly from the beginning of 2021. Other, short-term patterns are observed, with sometimes relatively large increases or decreases from one week to the next. This may mean that the underlying true behaviour of posting videos is rather variable at weekly level, or that there is some data processing going on at Youtube that we are unaware of. In any case, this plot suggests that shorter time periods than weekly are not relevant because of the high variance. Perhaps monthly numbers could be considered, as these will vary less and hence provide a more relevant series. Monthly data can be obtained from Youtube directly, or can be computed from the weekly series given in the figure.



Figure 27**.** Weekly number of ‘phone repair’ videos posted on Youtube from early 2017 until mid-2021.

#### E-commerce of refurbished electronic items

We use the number of reviews of each refurbished item as a proxy for the demand of this item: the assumption here is that only actual buyers of a refurbished item write reviews. The extent to which this assumption is valid is hard to verify. Nevertheless, bol.com can easily check that the account of the reviewer bought the reviewed item. In any case, the number of reviews is a lower bound on the actual demand.

We visualize the number of reviews of each item relative to the first observation date (October 27th, 2021) using the formula:

In this way the first observation always has a value of 100. Figure 28 shows that the number of reviews stays constant for the large majority of refurbished smartphones over the short observed time period. Only for one refurbished smartphone product the number of reviews increases. This result is not totally unexpected given the observed short time period.

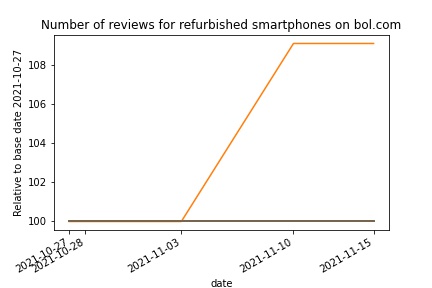


Figure 28 Number of reviews for refurbished smartphones on bol.com relative to base date 2021-10-27.

We found that there is only one refurbished laptop for which a review exists, and the number of reviews remained unchanged for this laptop. Therefore, we decided not to show the resulting figure.

#### Repair manuals (ifixit.com)

We focused on the view statistics of the 8 top-level category pages for this case study. The data contains view statistics of the last 24 hours, last 7 days, last 30 days and all time. Our main interest lied in deriving the overall interest in repair guides. Hence, the “all time” views statistic is probably most fit for our purpose. These view statistics vary wildly in size: from 531507 for Desktop PCs to 15559034 for Apple Mac in the first day of observation. In order to avoid compressing the figure by an extensive y-axis and as we are mostly interested in time trends, we visualized each observation after rescaling according to the formula:

This ensures that all view statistics start from a value of 100 in the first day of observation (November 3th, 2021). Figure 29 visualizes the trends. As view statistics can only increase, there is an obvious clear upward trend visible for all page categories.

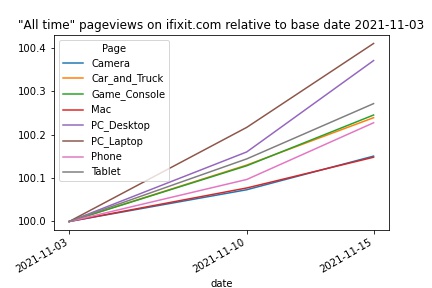


Figure 29 “All time” pageviews on ifixit.com relative to base date 2021-11-03.

### Summary, obstacles and outlook

Web scraping is a very broad topic and includes many potential applications other than the cases considered at present. The main benefit of the web scraping approach is that it can provide data that is (near) real time, and hence very up-to-date. Data can be scraped as soon as it appears online, and this can be done daily or even more frequently. Indicators developed from such data are very dynamic and up-to-date.

The extent to which there are specific other benefits or weaknesses of web scraped data depends on the specific websites that are used as a data source. Below we reflect on the data sources used in this study.

#### Repair videos

We conducted a first appraisal of repair videos as a data source and produced preliminary results shown in the previous paragraph. A major outstanding challenge is the development of one or more indicators that are relevant in the context of circular economy studies in general and of repair behavior in society in particular.

The basic search results as shown above and as can be obtained in other forms, does not in itself provide information that is immediately useful. Some strategy or approach should be developed that delivers a time series from which trends could be observed, of quantities that bear relevance to the topic of interest, in this case repair activities or behavior in circular economy.

Steps that could be taken towards the development of an indicator could include the design of a sampling strategy, for example the selection of a limited set of relevant videos to track through time, with a mechanism to update this set with new videos at set points in time. An appropriate data collection strategy should be devised that collects the necessary data at daily, weekly or other frequency.

One severe limitation is that the data that is available through the mechanisms discussed here is not disaggregated by country or smaller geographic regions. This data, however, is available at Youtube, since owners of published video materials can see such disaggregation. A route to explore could be to try to obtain access to this data, either via Youtube or via owners of certain Youtube channels. Privacy and data access rights come into play then, and care must be taken not to violate any of those.

There are more fundamental issues that deserve more attention too. First, the representativity of data obtained from Youtube: does this give a faithful representation of search behavior concerning repair instructions? There may be other sources, video platforms or other, that are consulted more commonly. If Youtube is only a minor source of information that is consulted by a specific population subgroup (socio-economic or age-based, for example), using its data could produce biased results.

Second, from a circular economy perspective, interest is in the extent to which people repair devices as opposed to buying new ones. From the repair video data alone, it is not possible to assess this, as it is not certain that people who look for and watch repair videos also actually repair their devices. Research addressing this would be needed and would entail the collection of additional data. It is not straightforward to develop a research study design to address this question. As long as this is not answered, results based on video data should not be proclaimed to reveal more than what they do: measuring online video posting and viewing behavior.

Third, approaches like that followed in the present work rely to a certain extent on Youtube and its internal operations and algorithms. For example, the results of a search algorithm come about in a way that is not publicly documented. When we search for videos on ‘phone repair’, how is it decided that a certain video is a hit or not? When we sort by relevance, what does this mean, exactly? These and similar issues are corporate information from Google and remain unknown. Even worse, they can change their algorithms at will and we would not know it – running the risk of observing changes in trends in our indicator time series that are caused by such algorithm changes rather than by true changes in society. Furthermore, we must assume that the data that is provided is faithful and not tampered with. There are no procedures to conduct any form of quality control or audit.

When appropriately framed and taking into account its weaknesses and concerns, repair video data does provide a quick, dynamic data source that is fairly easy to obtain. The development of an indicator from this data deserves more attention and is not without risk. Perhaps in combination with other, related indicators, it could have its place in a larger indicator set.

#### E-commerce of refurbished electronic items

We did a first analysis and assessment of using product reviews from an online retailer as a data source to derive an indicator that measures the demand for or interest in refurbished products. As actual sales data is not publicly available we resorted to the number of product reviews as a proxy measure. A further challenge is the construction of an indicator that reflects demand changes over time for refurbished products. The raw data we collected and presented here in itself is not a suitable indicator as the number of reviews for a product monotonically increases over time.

The data is directly obtained from the bol.com catalog and as such presents an up-to-date snapshot of the products sold. Collecting this data over time can then provide insight into the evolution of the refurbished product catalog. Currently the product catalog of bol.com for refurbished products is quite limited. As this catalog grows in the future, a sampling strategy needs to be implemented that tracks a basket of products which are then used to construct the indicator. The sampling strategy also needs to consider and account for products that disappear from the catalogue.

A major limitation lies in usage of the number of reviews for a product as a proxy for customer demand. First, we cannot check that each product review was written by a customer who actually bought the refurbished product. However, bol.com would be able to check this and it might be possible to verify with them what checks they have in place to ensure this (if any).

Second, there might be considerable lag between a customer buying a refurbished product and the customer writing a review. Customers have the right to return a product within 30 days of purchase and likely first use the product for some time before writing a review. Thus, we do not know the delay between product purchase and product review. Again, it might be worthwhile to reach out to bol.com to inquire if they have more insight into this delay. In this way one might be able to adjust the indicator for this lag.

Finally, the data obtained from bol.com might not be a representative sample of the whole population. While they have a large market share in The Netherlands and Belgium, there are other big players in the market that also sell refurbished products. Ideally this data needs to be combined with other data sources to construct a representative indicator.

#### Repair manuals (ifixit.com)

We have been collecting page view statistics from a popular website with repair guides for electronic devices in the hope that these might convey some information about the population's interest in repair for electronic devices. We found that the data is easily collectable, view statistics are available for individual devices (i.e., granular data) and the data has no missing values. Interpretation of the data is also straightforward.

This raw data can be used to construct an indicator that – at least partially - informs about the population's interest in self-repair of electronic devices.

However, some caveats and limitations do apply. First, the data stems from an entirely online data source and therefore is likely more representative of the internet-savvy part of the population. However, by combining this data with other offline data sources the resulting indicator could be made representative.

Second, a limitation of the current data study is that we only collected data from the 8 top-level page categories. The assumption here is that the majority of the site visitors start from this landing page and gradually click through the categories until they reach the desired product page. This is likely not realistic as it disregards the reality that many visitors probably reach their desired product page immediately using an external search engine. A major limitation of the current approach is that the collected view statistics of the 8 top-level categories do not capture these visitors. A sampling strategy could be implemented that tracks the individual view statistics for a basket of product pages instead of the current 8 top-level category pages. This would mitigate this scenario.

Finally, we must take the accuracy of the view statistics on faith as we cannot check their objectiveness. Although IFixit could manipulate these numbers, we judge this risk to be negligible.

*3.3.8.4. Overarching conclusion*

We can draw the following overarching conclusions from the case study on web scraping:

* Representativeness: each individual data source in itself is likely not representative for the overall population. I.e., Youtube views or page views on IFixit do not necessarily mean that people actually repair their devices nor does it mean that the general population has increased interest in repairing devices. Usage barriers might exist for certain parts of the population (i.e., the less technologically inclined) for both online data sources. Combining different data sources (e.g., from offline repair cafés) could alleviate this problem to some extent.
* Data objectiveness: the 3 data sources stem from commercial parties. Currently, we must take the data on faith (Youtube views, page views on IFixit, number of product reviews on Bol.com) with no verification possible. In case of Youtube and Bol.com, the algorithms behind the search API are to some extent black boxes (less so in case of Bol.com as they work with category ids and more so in case of Youtube).
* Sampling strategy for basket of tracked items is needed: the large volumes of available videos on Youtube, pages on IFixit and the expected increase in refurbished items on Bol.com highlight the need for a robust sampling strategy that allows tracking a basket of items for any derived indicator.

## Repair data (case studies)

### Introduction

Provoked by the circular economy philosophy, new forms of consumption (e.g. shared use, product-service systems, willingness to pay more for durability) have arisen. Existing indicators for measuring the circular economy uptake across European member states do not provide information on how consumption patterns are changing towards more circular products and services. With this case study, it has been investigated if we could find data streams that could allow for assessing evolutions in changed mindsets regarding repair.

As a first step, a literature review was carried out in order to compile existing national initiatives promoting consumers to choose product repair (and/or polices aimed for stimulating job creation but having more repair activities as a side effect). Having installed such a policy allows for tracking the application of it. In Sweden a tax deduction on (certain) repair activities applies and hence repair activities could be monitored. The Swedish data on tax deductions as reported by the Swedish Tax Agency was analysed more in depth. In parallel, another emerging data stream on repairs logged by repair communities was explored. The data is produced by citizens and/or volunteers repairing broken devices as brought to repair cafés and events.

#### Policy context

On the EU level, a few policies apply for repair. Most prominent examples include:

* The **Waste Framework Directive** (2008/98/EC) promotes waste reduction and defines different waste management concepts, including recycling and reuse. By setting a waste hierarchy, EU Member States are encouraged to prevent initial generation of waste, though at its occurrence, to prioritise repair and reuse of products, followed by recycling, recovery and lastly, disposal;
* The **Ecodesign Directive** (2009/125/EC) encourages improvement of environmental performance of electronic products. The main focus of the Directive is energy efficiency; however, it can be used to regulate other aspects of the product life cycle. To incentivize consumers to choose product reparation, products need to meet a certain quality standard and potentially engage some level of sentimental value. On October 1st, 2019, ten new rules regarding ecodesign for certain products were accepted by the European Commission[[8]](#footnote-9). These rules must be implemented by Member States by 2021, and include regulations regarding the availability of spare parts, stating that they should remain available from the producer up to a decade after purchasing a product;
* The main objective of the new **WEEE Directive** (2012/19/EU) is to promote reuse, recycling and recovery of electric equipment. The Directive sets criteria on the collection, treatment and recovery of e-waste;
* The **Consumer Sales and Guarantees Directive** (1999/44/EC) clarifies that any product bought within the European Union will be covered by a consumer guarantee for 2 years. During this time, the producer must ensure that the product is usable and functioning. However, after 6 months the consumer must personally prove that the product was defect at the time of purchase and not due to mishandling (Deloitte, 2016). In relation to repair, this policy increases the reliability of buying repaired goods, hence encouraging consumers to buy repaired products;
* Furthermore, the EU commission adopted a **Circular Economy Action Plan** in March 2020. A large set of measures in particular focussing on the electronics sector in Europe, will be implemented. Among others, the commission will be establishing the “right to repair” and a legislative proposal for a sustainable product policy initiative (including digital product passports).

Apart from those initiatives, each European Member State is free to implement repair related measures that extend beyond the existing Directives. A few Member States appeared to have come furthest with implementing fiscal policies influencing consumer repair choices. This type of policy has the advantage that it creates an accompanying data stream. Table 7 describes a few examples of fiscal policies at national level aimed for stimulating repair activities and simultaneously creating a new data stream.

But there are also examples of other, non-fiscal, national policies oriented towards stimulating repairs and simultaneously creating a data stream. A few examples are described in Table 8.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **Since** | **Name** | **More clarification** | **Related data stream** |
| Sweden | 2009 | Taks reduction of 30-50 per cent for the labour costs | Repair and maintenance as well as conversions, laundry and extensions are counted, provided that such work is carried out in close connection with a dwelling that the client owns and in which he or she lives.  Maximum of SEK 75,000 per person per year. | Data is reported with details on type of work performed, working hours, cost of material, other costs, buyer information, geographical location, and amount of tax deduction.  The contractors having delivered the services are responsible for reporting the data to the Swedish Tax Agency. The data is publicly accessible for free. |
| Graz (AT)  (AT from 2022) | 2017 | Repair bonus (vouchers): consumers are reimbursed for up to 50 per cent of the repair costs, capped to 100 EUR per repair case and to 1200 EUR per year | Both commercial repairers and community repair initiatives can apply, though they should be registered in order to ensure the quality of the repair.  From 2022 on, all Austrians are to receive up to 200 EUR for the repair of electrical and electronic equipment financed by the Covid-19 recovery fund. | A repair register: list of companies and repair initiatives fulfilling repairs.  The data could report on:   * The amount of money spent on repair activities eligible for tax deduction; * Number of applications, and hence repairs achieved by commercial repairers and by community repair initiatives. |
| Thuringia (DE) | 2021 | Repair bonus (vouchers): consumers are reimbursed max. 100 EUR of the repair costs | Most consumers had their appliances repaired by specialised retailers or electricians (54%). About 20 percent of the repairs were carried out in independent repairs. The voucher was also used for repairs in electronics stores (12%) or with the manufacturer’s customer service (15%). |
| France | 2022 | Repair fund | The fund is financed via an eco-contribution paid by producers. Consumers who goes to a repairer, which is certified (and hence committed to fulfill quality criteria and investments in shared tools facilitating online diagnostics), is entitled to a reduction in the total price of the repair bill. The fund reimburses repairers for the discounts. | A network of certified professionals.  The data could report on:   * The amount of money spent on repair activities; * Number of applications, and hence repairs achieved by certified repairers. |

Table 7 Examples of fiscal policies at national level for stimulating repair and simultaneously creating a new data stream

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country | Since | Name | More clarification | Related data stream |
| France | 2019-2020 | Repair index | The French government adopted a law regulating the mandatory display of clear information for consumers on the repairability of electrical and electronic equipment. It applies to 5 categories of products sold in France after 1 January 2021 being: smartphones, laptops, televisions, washing machines and lawnmowers. (And there are more to come.)  <https://www.ecologie.gouv.fr/indice-reparabilite> | Each criterion being documentation, disassembly, availability of spare parts, price of spare parts and finally some more product-specific aspects, is scored on 20 points and aggregated to make the final grade.  Sellers are obliged to display the index near the point of sale. The manufacturer is obliged to make the index available to anyone who requests it. The index has to be displayed near the product in shops, and online next to the price of the product. |
| Finland |  | Variable warranty period | Products purchased in Finland do not come with a specified number of years for a warranty, regardless of whether it is a primary production or second-hand product. The length of the warranty is instead decided based on the predicted lifespan of the product. | Data on warranty claims containing useful information about product quality and reliability, though non publicly available |
| Norway |  | Extended warranty (5 years) | The warranty period was prolonged to 5 years. Additionally, people can choose to have the product repaired rather than replaced. |

Table 8 Examples of other policies at national level for stimulating repair and simultaneously creating a new data stream

### Data collection

#### E-governmental data

In Sweden, repairers can claim back part of the labour costs related to the repair. Two types of activities performed in dwellings are looked at, respectively RUT (covering maintenance, laundry and cleaning) and ROT (covering repairs, conversion and extension). It is up to the service provider to apply for the client’s preliminary ROT and RUT tax deduction and request a payout from the Swedish Tax Agency for that amount. The client only pays the remaining amount. This way, the Swedish Tax Agency is collecting the data[[9]](#footnote-10) on repairs automatically from applications for tax deduction which are submitted.

#### Citizen based data

The Open Repair Alliance or in brief ORA is a repair network covering a large set of repair communities active within several countries worldwide (Figure 4). The ORA consists of the following actors:

* Anstiftung Foundation: Research institution based in Germany that coordinates a repair initiative of 700 repair cafés across Germany;
* Fixit Clinic: American-based organisation that teaches in repair competencies by e.g., catching pop-up repair cafés;
* IFixit: A private company selling spare parts for repair and providing a wiki-based platform with more than 30 000 repair guides;
* The Repair Cafe Foundation: A network of 1 300 repair cafes worldwide collecting data on repairs from these;
* The Restart Project: London-based community planning repair events.

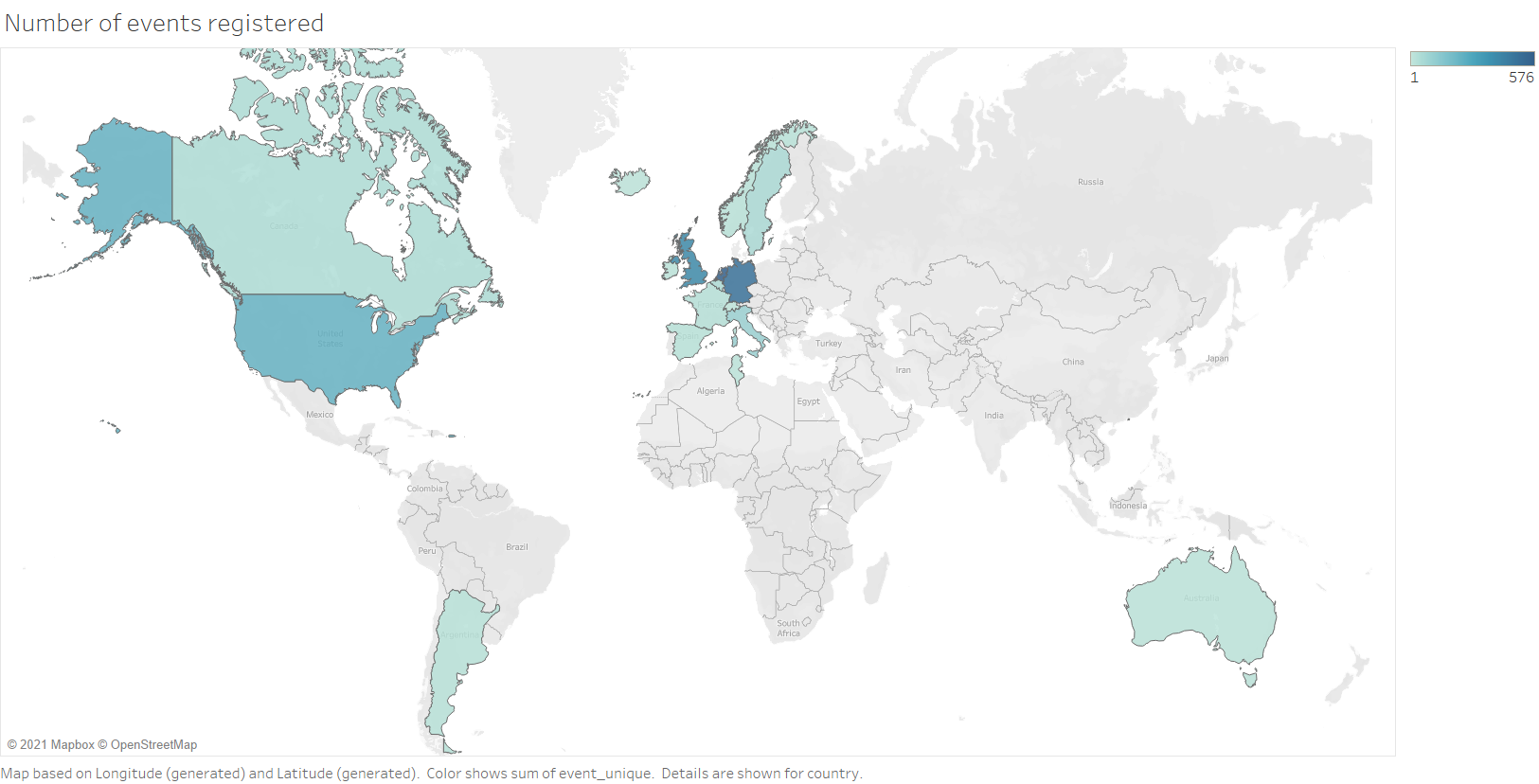


Figure 30 Number of events registered across countries based on ORA data

Driven by the scale effects of pooling the data generated at the repair cafés, they have developed a joint approach to document successes and challenges in repairing items and push their members making use of it. The ORA even have built an open data standard[[10]](#footnote-11) for the collection and sharing of open data on electronics repair. All repair networks can freely engage and start share their data. The data is aggregated and reported on a yearly basis.

### Data documentation and quality assurance

Each data set obtained and further analysed is described in a separate profile below.

|  |  |  |  |
| --- | --- | --- | --- |
| **E-governmental data** | | | |
| **Data set description** | **Source of origin:**  Swedish tax agency | **Data retrieval process:**  Download of (structured) data (5 tables) and metadata sheet via API  [API's en open data | Belasting (skatteverket.se)](https://www7.skatteverket.se/portal/apier-och-oppna-data/utvecklarportalen/oppetdata/Rot-%20och%20rutbetalningar,%20statistik) | **Storage format:** xlsx |
| **Accessing date:** 3th August 2021 | **Geographic scope:** Sweden | **Temporal scope:** 2009-2021 |
| **Level of aggregation:**  Monthly tax deductions rewarded (number and value) divided between type of work performed (ROT and RUT), being explicit about number of buyers and executors and providing geographical locations (regions and municipalities). | **Features:**  N/A | **Volume:**  file size: 667 KB |
| **Quality assessment** | **Trustworthiness:** The main challenge is the self-report method which brings some biases, since not all data is reported in the correct categories. However, as the data is reported by the companies carrying out repairs and not by consumers, coverage is likely to be comprehensive and broadly representative of activity within each category. | **Currentness:** last updated on 11th March 2021 | **Accuracy**: It is official data used and published by the Swedish Tax Authority. |
| **Completeness:** No missing data | **Representativeness:** Tax breaks on repairs are currently not widespread policy in EU MS, therefore representative data for the whole EU is unattainable. | **Interpretability:** Metadata provided on a separate sheet |

|  |  |  |  |
| --- | --- | --- | --- |
| **ORA data** | | | |
| **Data set description** | **Source of origin:** The Open Repair Alliance | **Data retrieval process:**  Download of (structured) data (5 tables) and metadata sheet via API | **Storage format:** CSV |
| **Accessing date:** 23th August 2021 | **Geographic scope:** So far data reported from 18 different countries worldwide | **Temporal scope:** 2012-2021 |
| **Level of aggregation:** Best reported are the product category and the repair status. If known, one could also provide the year of manufacturer and the brand. The failure and solution (if found) could also be described in more detail. | **Features:**  N/A | **Volume:**  file size: 6,7 MB |
| **Quality assessment** | **Trustworthiness:** All repair initiatives can freely apply the standard and share their data. The repair standard has been commonly agreed between the stakeholders of the alliance. The broad involvement and the transparent approach make the standard of measurement trustworthy. Some caution should be exercised with regard to the interpretation of reported repair volumes or success rates, as there is an interest of the data providing organisations (from individual repair cafés through to networks and alliances) to report high numbers in order to underline their importance and to push policy. In the case of repair café locations, however, this might be less the case as they are more verifiable. | **Currentness:** last updated on 22nd February 2021 | **Accuracy**: A certain degree of self-selection bias applies as repair café visitors are unlikely to be representative of the broader population, and subsequently the products they bring to the event may also not be representative of the broader market. |
| **Completeness:** rather weak | **Representativeness:** The data inform about products that consumers are seeking to repair at a repair café or a similar facility. The metrics do not include repairs taken place at home, or by professional repairers. The network does not cover all MS yet, but the standardisation offers the possibility of comparatively easy dissemination. | **Interpretability:** rather weak (lots of outliers) |

### Data integration

For both data sets no further integration step was needed. But the data offered by the Open repair Alliance is resulted from a large data integration process, carried out by the Alliance on an annual basis.

### Data preparation

As both datasets are structured in tables, a reshaping of the data was needed in order to be able to analyse them. The citizen driven dataset does contain a lot of outliers and gaps. Outliers had to be removed before analyzing. The data are organized around a large set of product categories. This list is created by themselves and hence focused upon products that are typically brought to repair events (e.g. large and heavy appliances such as washing machines are missing). It would have been better if the list of product categories had been aligned with more EU official categories such as CN classification or PRODCOM.

### Data analysis

#### E-governmental data

The data stream can provide to what extent repair activities are taking place in the country and per region and how this is evolving over time. The data is also explicit about the number of executors (asking for the tax deduction) and the corresponding amount of “buyers”, this is the clients they offered services to.

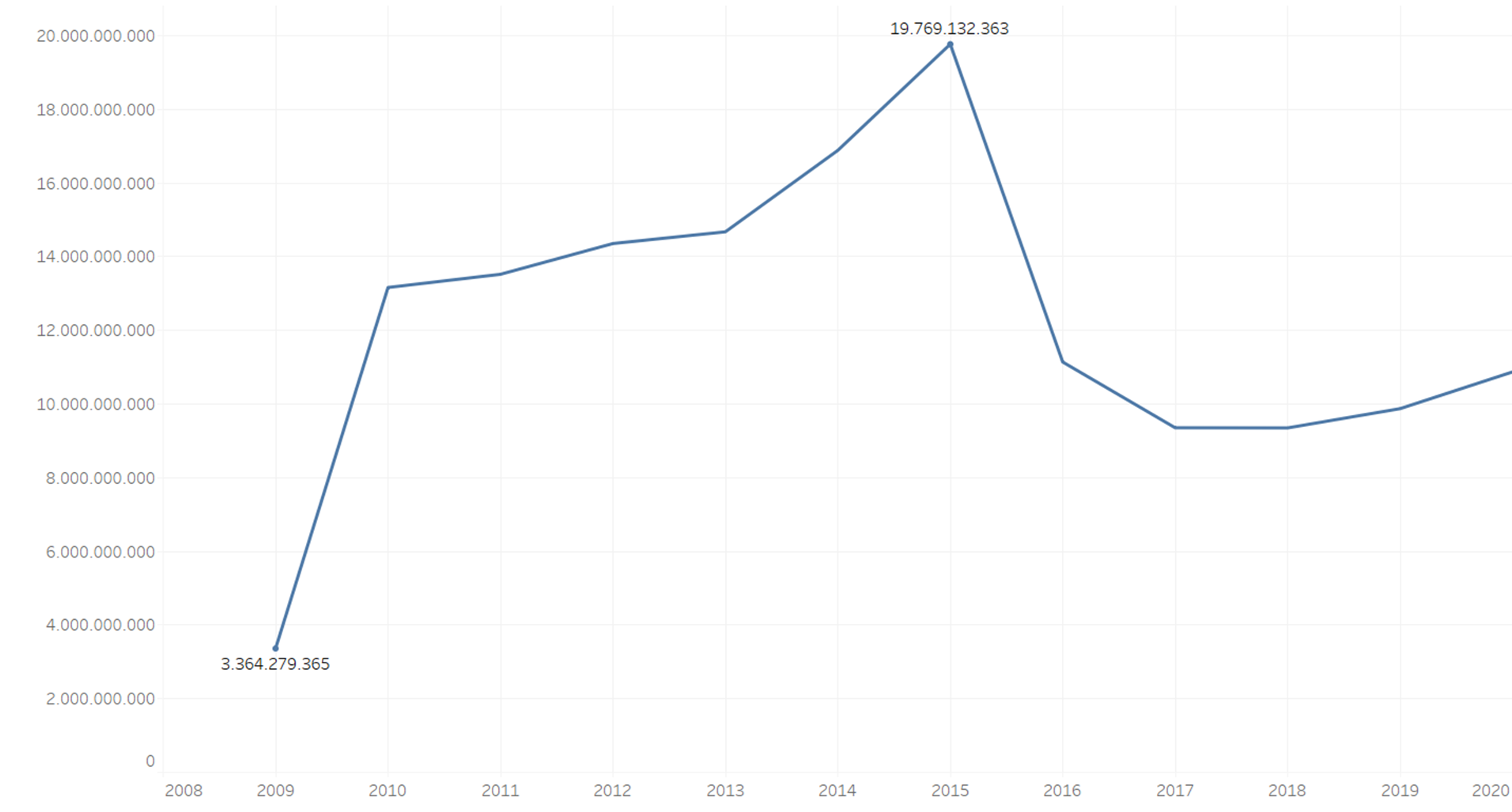


Figure 31 Total amounts of repair deductions registered in SEK since the start in 2009

The overall financial support for repair activities has been increasing steadily from 2009 until 2015 to almost 20 billion SEK, which corresponds to roughly 2 billion EUR. Though in 2016 a considerable smaller amount of support (11 billion SEK) was provided. Although over the last years the numbers are again increasing, it has never reached the 2015 levels. Looking at the Swedish repair sector[[11]](#footnote-12), both, the number of enterprises and persons employed have been growing steadily since 2009 in (source: Eurostat, Annual detailed enterprise statistics for services in Sweden).

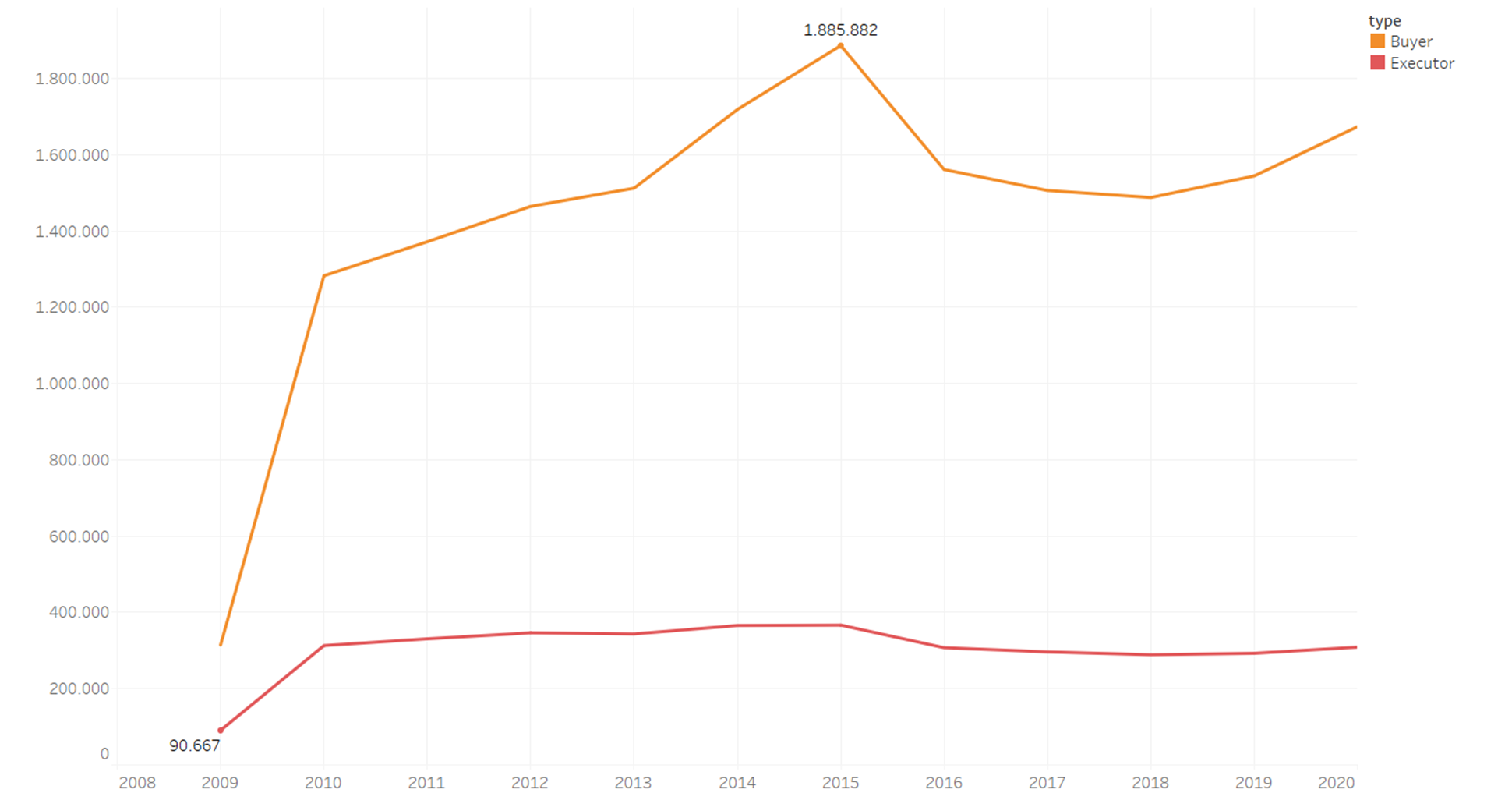


Figure 32 Number of buyers and executors of repairs since the start in 2009

The Swedish Tax agency also report the number of executors and buyers of repairs. The executors, considered as the contractors having delivered the services and responsible for reporting the data, are rather stable over the years (+/- 323 606). In 2020, the repair executors have received on average 3646 EUR for carrying out repair activities. Each executor undertakes yearly on average 5 to 5,5 submissions. Also here, a decrease between 2015 and 2016 can be observed. In 2016, 16% less executors of repairs were registered.

The number of buyers of repair services undergoes a similar trend. After 2015, a drop of 17% of citizens purchasing repair services is observed. Although in 2020 up to 991 002 citizens have paid substantially less for their repairs due to the tax reduction rewarded by the Swedish government. In 2020 each buyer received on average 665 EUR.

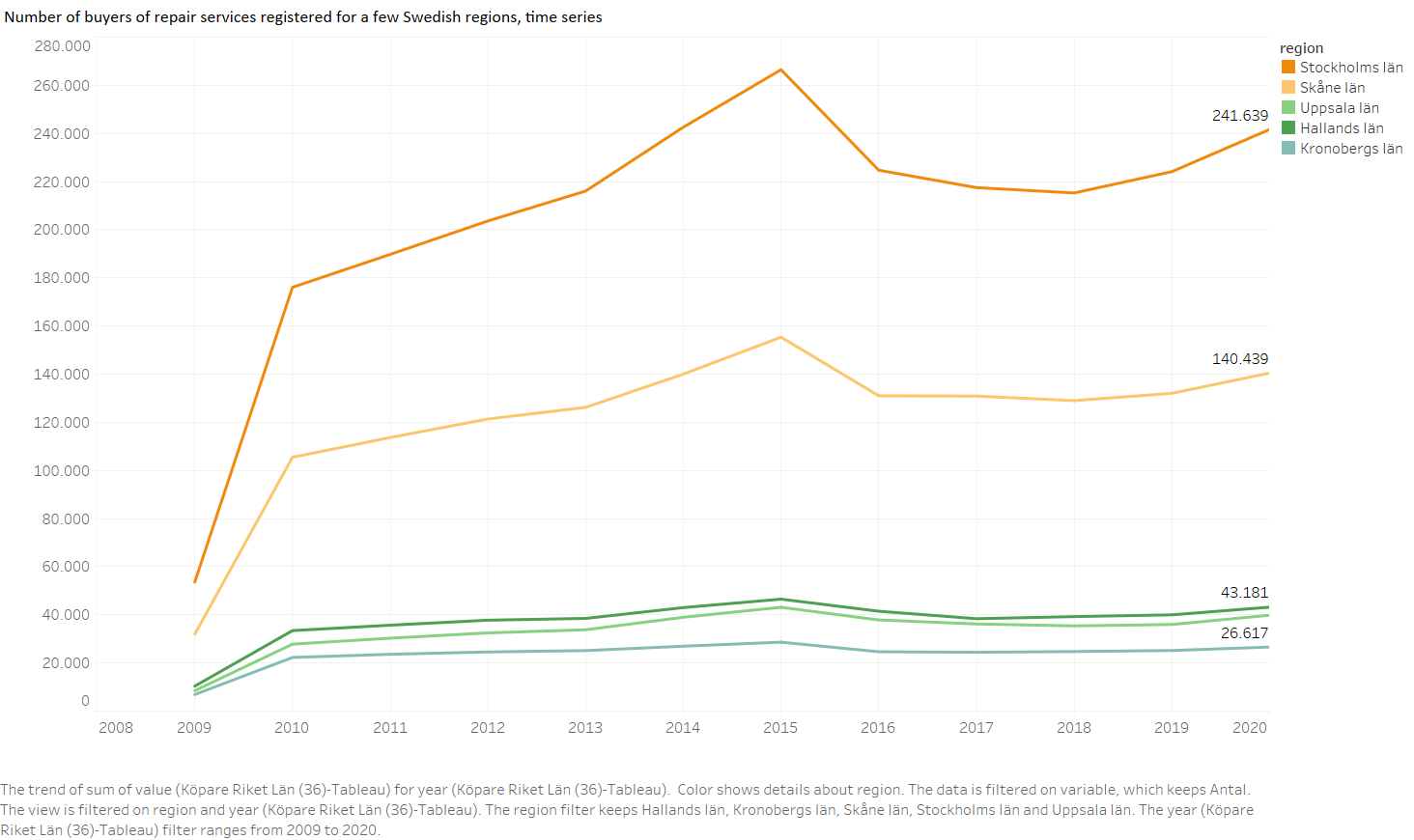


Figure 33 Number of buyers of repair services registered for a few Swedish regions since the start in 2009

Zooming in on geographical differences, Figure 33shows a similar trend in number of buyers of repair services across a set of Swedish regions.

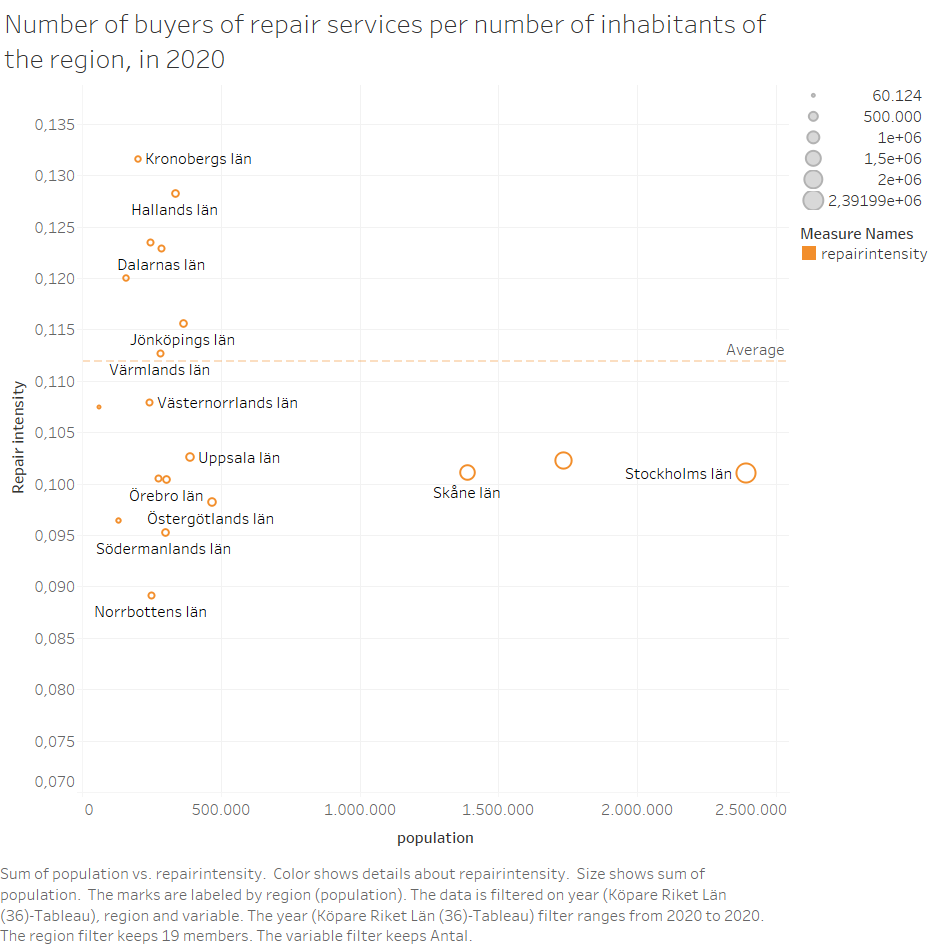


Figure 34 Number of buyers of repair services per number of inhabitants for the Swedish regions, compared against the Swedish average (dotted line) in 2020

A certain level of dissimilarity across the Swedish regions exist with respect to repair attitudes. Figure 34 sets out the repair intensity. This is calculated by dividing the number of buyers of repairs in 2020 in a region with the number of inhabitants of the region in 2020. The larger and more densely populated Stockholm region is performing worse than a few more thinly populated regions such as Hallands and Kronobergs.

#### Citizen driven data

Although strongly growing over time, the repair network and hence the repair events covered, is still limited. Many more repair networks, whether official or more occasional, could be attracted. So far, a growth in the number of repairs can largely be seen as a growth in the network. In this early stage it cannot be used for monitoring a behavioural change of the citizens. If further adopted by repair communities worldwide, the data stream could be used for evaluating how much repair is taking place. If sufficiently covered, the data stream could even be used for identifying geological differences across countries, regions, or between rural and urban areas[[12]](#footnote-13). Though the data stream as it is today, could be used for:

* Monitoring the **number of events and visitors** for a city or region;
* Monitoring the **success ratio** of repair (overall and per product category);
* Identifying repair **barriers** per product category;
* Identifying **common failures** per product type (and differences between brands);
* Monitoring **incoming products**: total, per brand, per product category and their age.

The data stream allow to calculate and monitor the number of successful repairs as a share of all attempted repairs. Table 9 shows the average number of incoming products per event and average success rate since 2012. While started very marginal, the number of incoming products as logged by the system drastically increased and hence, the success rate as achieved becomes more reliable over time.



Table 9 Average number of incoming products per event and average success rate, time series, based on ORA data (total)

The data stream allow to derive the repair status of the devices per product category. There are three possible repair statuses:

* Fixed: the failed product is repaired during the Repair Café;
* Repairable: the product is assumed to be repairable but the repair cannot be finalised during the Repair Café (hence the process was considered uncomplete);
* End of life: the product is assumed to be unrepairable.

As presented by Figure 35 repair success rates vary across product categories. Roughly half of the repairs are successful today. The data provide more details on the barriers the repairers have been faced with. Examples include: spare parts that are not available or available but too expensive; no way to open the product; repair information cannot be retrieved; lack of equipment; product too worn out. Looking at mobile phones only, multiple fault types are involved, among which 41 per cent were linked to screens, 15 per cent to battery and other power-related issues; 6 per cent were linked to charging ports.

If the failure has been identified by the repairer, this information can also be uploaded in the system. This way, the dataset can be used for retrieving common faults such as irreplaceable batteries. In turn, this information might give an indication on how robust product designs are. Although, similar information can also be found through laboratory studies that might provide data of higher quality.

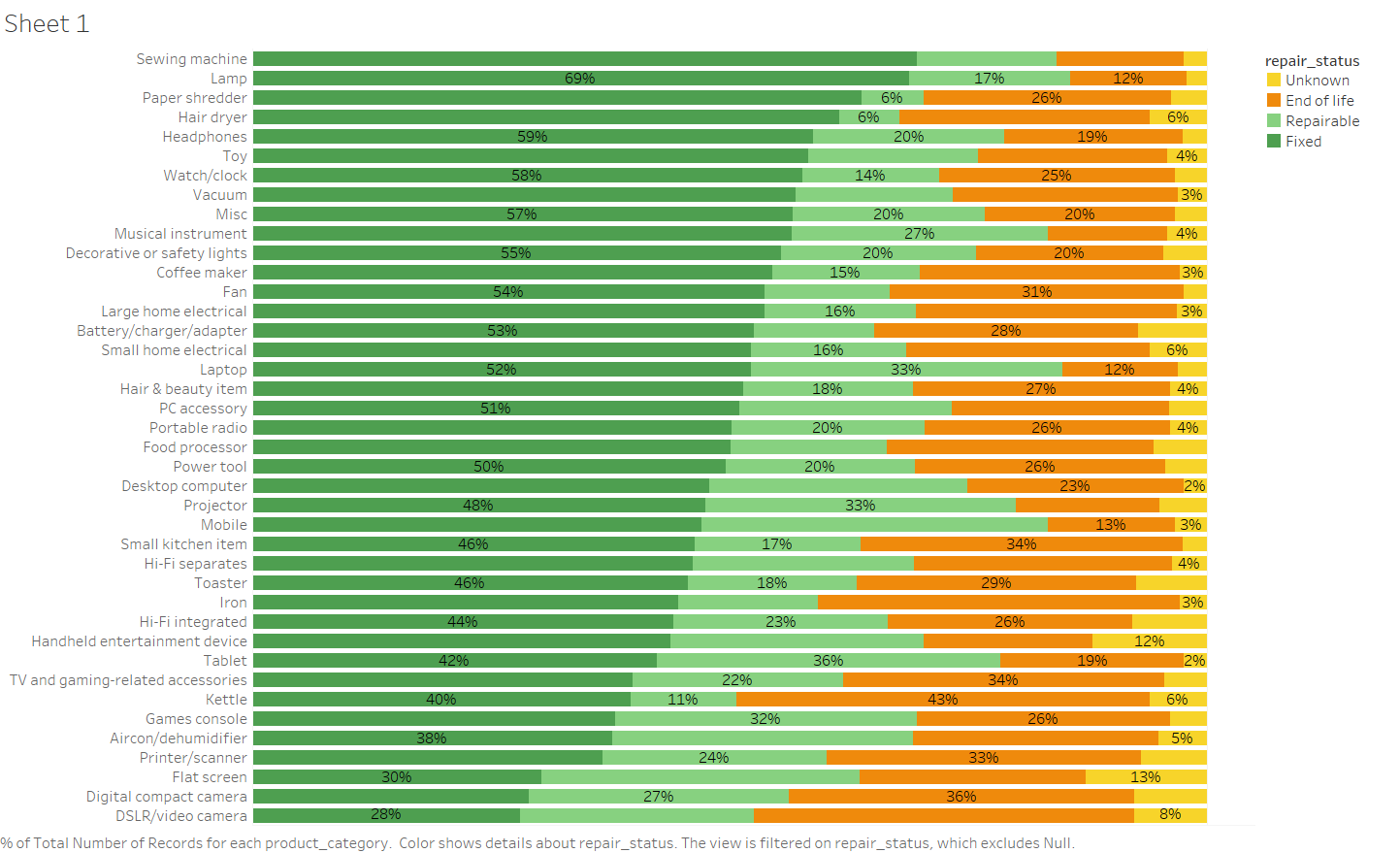


Figure 35 Repair status (total) per product category, based on ORA data

In case the manufacturing year of the device is known, the data stream is also revealing the age of incoming products. As Figure 36 is indicating, a large part (45%) of the devices brought to these events are younger than 6 years old. Looking at smartphones only, 44 per cent of the failed devices are 2 years old or younger.

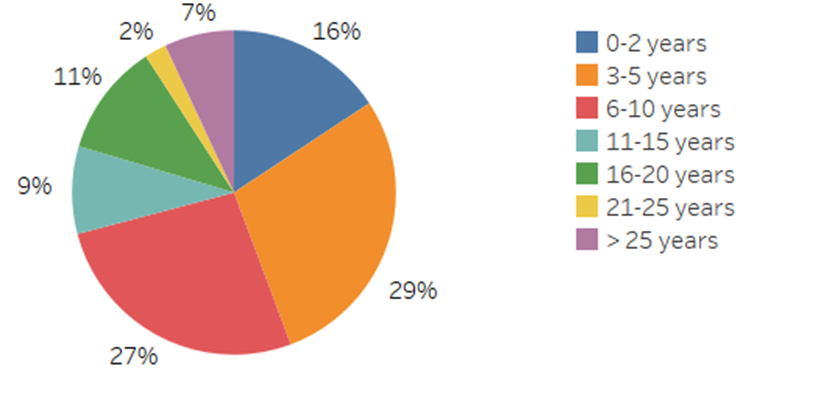


Figure 36 Age of incoming products per category, based on ORA data (total)

The circular economy promotes longer product lifetimes, which is in practice hard to monitor. Though an indication of product lifespans can be offered by the data stream since average ages of devices considered as end of live can be derived. This could serve as a proxy for lifespans of EEA categories. Figure 37 shows the number of years the device has reached before reaching its end of live by means of a boxplot for a set of product categories. In literature (EEA, 2019) desired lifespans can be found. The desired lifetime has been defined as the average time that consumers want products to last. If a devices is not able to reach the desired lifetime, the product owner might experience this as disappointing. If this happens to be the case for a certain product type or brand, a conjecture for premature obsolescence might come.

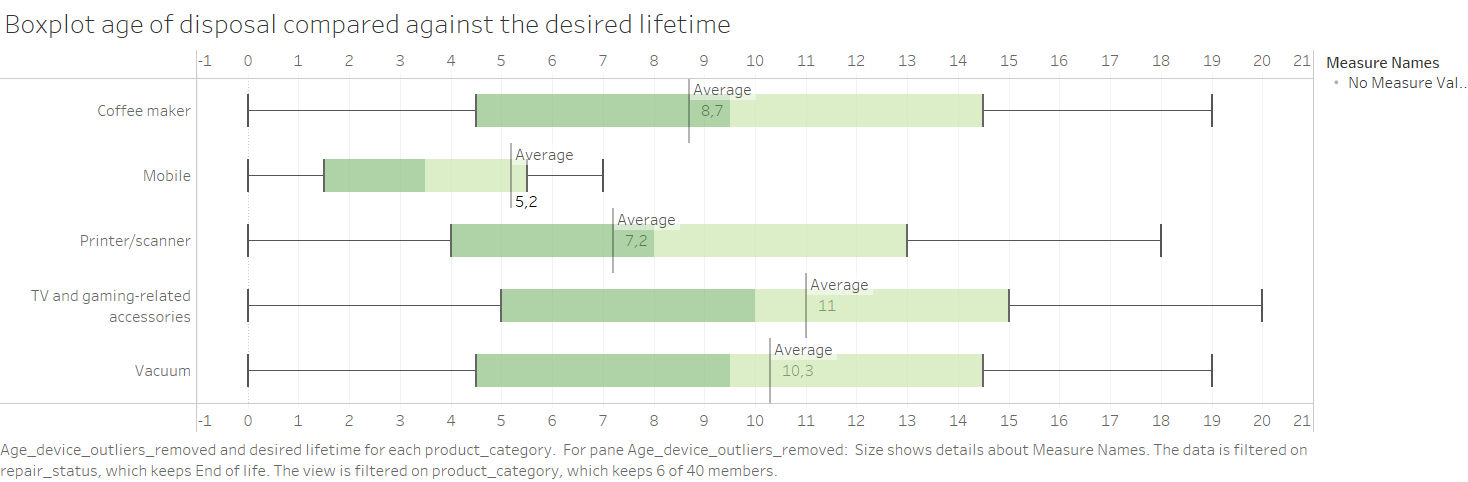


Figure 37 Age of disposal plotted for a set of product categories based on ORA data and compared against desired lifetimes (derived from literature)

### Summary, obstacles and outlook

Both datasets emerged from different processes and hence shed a different light on repair considered as one of the value retention strategies. They both bring value and might play a role in nourishing a CE indicator showing how consumption patterns are changing. However, both data streams come with a few limitations the user should be aware of.

#### E-governmental data

When it comes to e-governmental data, those are most relevant and collected, processed and offered by administrations what makes them credible and reliable. Though they only arise in a few member states having installed a policy for stimulating repair of a (limited) set of product categories. So far the data cannot be used to evaluate the repair behavior of citizens across Europe.

Usually, policies are not static. They adapt towards different market circumstances or enhanced insights. Once a policy has changed, the accompanying data stream will represent the new structure, and a structural break in the time series might pop up. The quality of the time series strongly relies on changed policies. This hinders the development of a metric reflecting long term trends.

It depends upon how the policy is defined, but for most cases it is opportune that the producers, but sometimes also the consumers, are set responsible for asking the deduction. This structure does not ensure full coverage. There will always be producers or consumers not being aware or capable to successfully submit their application. Same goes for the quality of the data: although some checks can be done, the data will not be free of mistakes.

To further investigate the possibilities the data stream could offer, one could look for complementary data available (e.g. Eurostat, household budget surveys, input output data) and evaluate the relevance of a combined view.

#### Citizen driven data

Datasets generated by citizens themselves certainly come with quality limitations. But that does not mean that they are worthless. The more data can be pooled the better. That is why the rolling out of the data standard towards many more repair events and repair networks over Europe is important.

At these event, repairers offer their expertise and time on a voluntary basis. What means that the social aspect of the event is put forward. Citizens highly appreciate the interaction. Often the citizen is taught how the failure could be prevented, or how to solve themselves in case it happens another time. One could encourage the volunteers to fill out the data properly but it may not become a discouraging factor.

Repairs are not only realised at repair events. The commercial repairers are also conducting repairs. The involvement of commercial repair organisations which collect data in multiple different ways, certainly is important for improving the quality of the data stream. Lastly, some broken devices are repaired not by involving a commercial repairer, neither by bringing it to a repair event. This is called “self-repair” and it is hard to cover this activity fully. Although some indications could be gathered by tracing the availability of repair videos and manuals (see section 3.3). The more the data stream could cover repair data, the better. This process could be stimulated by the EC by encouraging citizens, repair networks and commercial repair to participate. Or via including repair data in the so called product passports, as currently developed.

# Conclusion and outlook

We report on specific hands-on explorations conducted with new data sources that could be used for circular economy monitoring. Earlier work consisted of identifying priority questions and listing potential data sources. The data sources considered in this report can be used to obtain information on the uptake of sharing systems, the adoption of value retention strategies and the degree to which consumers switch to circular products and services.

Since our explorations were the first of their kind, data sources were chosen that are easily – mostly publicly – available, trustworthy, representative and novel. Novel in this context means that we considered data sources that have not been used for our purposes yet. These sources are therefore not necessarily new, it is their use for circular economy monitoring that is new. This repurposing of data poses several challenges when using them for statistical purposes and policy assessments (Daas et al. 2015).

In this report we present a detailed appraisal of the following data sources:

* Browser fingerprints: mobile devices leave traces when using apps and visiting websites; these traces include information on the make, model and/or operating system of the device and can be used to provide information on life-time of devices.
* Google searches related to car sharing: these can reveal trends in interest in car sharing; in our work these are assessed in comparison with detailed publicly available data on car sharing use in Germany.
* Web scraping of repair instruction data and e-commerce of repaired/refurbished electronics: information that is available on websites can automatically be collected - '*scraped’* - and turned into data that can be used for analysis; we considered popularity of repair videos, repair manuals, and offerings of repaired or refurbished items in online web shops.
* Collecting diverse repair data sets: a variety of patchy data sets is available online and can be brought together to provide a comprehensive view on repair initiatives and uptake and adoption of circular consumption patterns. Data sets provided by governmental and non-profit organizations are studied.

The gains to be had from using existing data streams for new purposes include:

* Currentness: many of the data sets considered are available in near real-time when or shortly after the events happened. This is in contrast with official statistics which are typically available only a while after a certain reporting period such as a year or quarter.
* Frequency: many of the new data sources can be obtained or collected at a much higher frequency than official statistics. Weekly and sometimes daily data updates are possible.
* Data collection costs: the data considered in this report were typically collected for other purposes and came about for that purpose or as a side effect of it. When such data is repurposed, such as for circular economy monitoring in our case, there is no cost associated with data collection, in contrast for example with setting up and executing a dedicated survey.

While the potential gains are appealing and promising, none of the data sources we considered can be used as-is, without any further work. When these quantitative data are used for policy making and monitoring, a number of challenges need to be addressed:

* Representativity: while a sample survey is designed to obtain a representative sample from a certain target population, the new data streams are not designed that way. They typically cover subpopulations that use a particular service, product, or website. Hence care must be taken when inferring population level attributes from these new data sources; selective data sets can result in biased estimates.
* Transparency: since data are collected, processed, and provided by other parties, the exact underlying procedure or mechanism is often not fully known. This may hamper explainability and transparency of the entire data life cycle.
* Control: since other parties are in control of the data generation, they could make changes to or even discontinue the data provisioning. This is a risk when setting up long-term monitoring schemes.
* Usability: the new data streams are often not directly informative. They often need transformations of various kinds, or must be incorporated into some data retrieval scenario for the data to be useful for indicator development in particular, and broader uptake in general.

Despite these challenges, we see future opportunities for research and development exploiting the benefits of these data sources. While each source has shortcomings and limitations, a promising direction would be to explore the joint use of multiple, independent sources. A composite indicator or joint analysis would be more robust against sudden errors or deviations in single sources and would be more representative. When constructing such composite indicators, care needs to be taken to obtain transparent and interpretable indicator definitions. Given this condition, these new data sources and derived indicators could provide insights into circular economy activities that remain invisible for monitoring otherwise.

# References

ADAC. (2021, February 18). Carsharing-Statistik: Corona-Krise bremst 2020 das Wachstum. <https://www.adac.de/news/carsharing-statistik-2020/>

Agora Verkehrswende. (2019). Bikesharing im Blickpunkt—Eine datengestützte Analyse von Fahrradverleihsystemen in Berlin. <https://www.agora-verkehrswende.de/fileadmin/Projekte/2018/Stationslose_Bikesharing_Systeme/Bikesharing_im_Blickpunkt_20190304_web.pdf>

AmIUnique. (n.d.). *What is browser fingerprinting?* Retrieved 3 November 2021, from <https://www.amiunique.org/faq>

Autohaus. (2020, June 15). *Zahl der Carsharing-Nutzer rückläufig: Boom mit Knick*. autohaus.de. <https://www.autohaus.de/nachrichten/autohersteller/zahl-der-carsharing-nutzer-ruecklaeufig-boom-mit-knick-2701966>

BCS. (n.d.-a). Aktuelle Zahlen und Fakten zum CarSharing in Deutschland. bcs Bundesverband CarSharing e.V. Retrieved 8 July 2021, from <https://carsharing.de/alles-ueber-carsharing/carsharing-zahlen/aktuelle-zahlen-fakten-zum-carsharing-deutschland>

BCS. (n.d.-b). CarSharing-Städteranking 2019. bcs Bundesverband CarSharing e.V. Retrieved 12 July 2021, from <https://carsharing.de/alles-ueber-carsharing/carsharing-zahlen/carsharing-staedteranking-2019>

BCS. (2021a). Die größten CarSharing-Anbieter gesamt (nach Flottengröße). bcs Bundesverband CarSharing e.V. <https://carsharing.de/presse/fotos/zahlen-daten/groessten-carsharing-anbieter-gesamt-nach-flottengroesse>

BCS. (2021b). Mitglieder. bcs Bundesverband CarSharing e.V. <https://carsharing.de/verband/wir-ueber-uns/mitglieder>

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113. <https://doi.org/10.1016/j.techfore.2017.07.027>

Daas, P. J., Buelens, B., van den Hurk, P. A., & Puts, M. J. (2015). Big Data as a Source for Official Statistics. Journal of Official Statistics, 31(2), 249-262.

Dalhammar, C. (Ed.), Richter, J. L. (Ed.), Almén, J., Anehagen, M., Enström, E., Hartman, C., Jonsson, C., Lindbladh, F., & Ohlsson, J. (2020). Promoting the Repair Sector in Sweden.

Deloitte, 2016, “Study on Socioeconomic impacts of increased reparability – Final Report”, Prepared for the European Commission, DG ENV.

Desjardins, J. (2018). This Chart Reveals Google’s True Dominance Over the Web. Visual Capitalist. <https://www.visualcapitalist.com/this-chart-reveals-googles-true-dominance-over-the-web/>

Deutsche Bahn AG. (n.d.) Flinkster—Open-Data-Portal – Deutsche Bahn Datenportal ABN. Retrieved 12 July 2021, from <https://data.deutschebahn.com/dataset/data-flinkster.html>

EEA, 2019, “Electronics and obsolescence in a circular economy”, from the EEA’s European Topic Centre on Waste and Materials in a Green Economy

EPA Network, EEA, ISPRA, & SNPA. (2021). *Bellagio Declaration – Circular Economy Monitoring Principles*. <https://epanet.eea.europa.eu/reports-letters/reports-and-letters/bellagio-declaration.pdf/view>

Google. (n.d.-a). Compare Trends search terms—Trends Help. Retrieved 15 July 2021, from <https://support.google.com/trends/answer/4359550?hl=en&ref_topic=4365530>

Google. (n.d.-b). Search tips for Trends—Trends Help. Retrieved 15 July 2021, from <https://support.google.com/trends/answer/4359582?hl=en&ref_topic=4365530>

Google. (n.d.-c). FAQ about Google Trends data. Retrieved 27 October 2021, from <https://support.google.com/trends/answer/4365533?hl=en>

Google Trends. (n.d.-a). Retrieved 8 July 2021, from <https://trends.google.com/>

Google Trends. (n.d.-b). Retrieved 11 August 2021, from <https://trends.google.com/trends/explore?cat=47&date=2004-01-01%202021-07-15&geo=DE&q=%2Fm%2F03t80x>

Google Trends. (n.d.-c). Retrieved 11 August 2021, from <https://trends.google.com/trends/explore?date=all&geo=DE&q=carsharing,%22car%20sharing%22,carsharing%20%2B%20%22car%20sharing%22,%2Fm%2F03t80x>

Google Trends. (2021a). <https://trends.google.com/trends/explore?date=all&geo=DE&q=%2Fm%2F0gh8bl8,%22cambio%20carsharing%22>

Google Trends. (2021b). <https://trends.google.com/trends/explore?date=all&geo=DE&q=car2go%20%2B%20drivenow%20%2B%20%22share%20now%22%20%2B%20%E2%80%9Cmiles%20carsharing%E2%80%9D%20%2B%20%E2%80%9Csixt%20share%E2%80%9D%20%2B%20weshare,greenwheels%20%2B%20%E2%80%9Cstattauto%20m%C3%BCnchen%E2%80%9D%20%2B%20flinkster,stadtmobil%20%2B%20%E2%80%9Ccambio%20carsharing%E2%80%9D%20%2B%20teilauto%20%2B%20book-n-drive%20%2B%20%E2%80%9Cbook%20n%20drive%E2%80%9D,carsharing%20%2B%20%22car%20sharing%22>

Google Trends. (2021c). <https://trends.google.com/trends/explore?date=all&geo=DE&q=car2go%20%2B%20%22share%20now%22,car2go%20%2B%20drivenow%20%2B%20%22share%20now%22,%2Fm%2F0bmj5jx>

Gschwenter, M. (2021). *Browser-Fingerprinting—Was ist es und wie kann ich mich schützen?* <https://www.experte.de/it-sicherheit/browser-fingerprint>

Janc, A., & Zalewski, M. (n.d.). *Technical analysis of client identification mechanisms—The Chromium Projects*. Retrieved 3 November 2021, from <https://sites.google.com/a/chromium.org/dev/Home/chromium-security/client-identification-mechanisms#TOC-Machine-specific-characteristics>

Keenan, J. (2021). *What is Browser Fingerprinting?* Smartproxy. <https://smartproxy.com/blog/what-is-browser-fingerprinting>

Kirchherr, J., Reike, D., & Hekkert, M. (2017). Conceptualizing the circular economy: An analysis of 114 definitions. *Resources, Conservation & Recycling*, *2017*(127), 221–232. <https://doi.org/10.1016/j.resconrec.2017.09.005>

Kobusińska, A., Pawluczuk, K., & Brzeziński, J. (2018). Big Data fingerprinting information analytics for sustainability. *Future Generation Computer Systems*, *86*, 1321–1337. <https://doi.org/10.1016/j.future.2017.12.061>

Mann, B. (2020, August 3). *Browser Fingerprinting [2020 Update]—What Is It & How to avoid It*. <https://blokt.com/guides/browser-fingerprinting>

Rohatgi, A. (2020). Webplotdigitizer: Version 4.4. <https://automeris.io/WebPlotDigitizer>

Statista. (2020). Größte Carsharing-Anbieter nach Kundenzahl 2020. Statista. <https://de.statista.com/statistik/daten/studie/259118/umfrage/carsharing-anbieter-nach-kundenzahl/>

Tillmann, H. (2013). *Browser Fingerprinting: Tracking ohne Spuren zu hinterlassen—Diplomarbeit zur Erlangung des akademischen Grades Diplominformatiker* [Humboldt-Universität zu Berlin]. <http://www.bfp.henning-tillmann.de/downloads/Henning%20Tillmann%20-%20Browser%20Fingerprinting.pdf>

Wikipedia. (2021a). Greenwheels. In Wikipedia. <https://de.wikipedia.org/w/index.php?title=Greenwheels&oldid=213714322>

Wikipedia. (2021b). Share Now. In Wikipedia. <https://de.wikipedia.org/w/index.php?title=Share_Now&oldid=215242829>

ANNEX: Meeting minutes

# Novel data types for CE monitoring: expert workshop

Wednesday October 27 2021, 10:00 - 12:00

Microsoft Teams Meeting

**Participants**

|  |  |
| --- | --- |
| **ETC Team** | |
| Peder Jensen | Accepted |
| Nora Brüggemann | Accepted |
| Yoko Dams | Accepted |
| Bart Buelens | Accepted |
| Pieter Jan Kerstens | Accepted |
| Maike Jansen | Accepted |
| János Sebestyen | Accepted |
| Holger Berg | Accepted |
| Vercalsteren An | Accepted |
| Lize Borms | Accepted |
|  | |
| **NFP** | |
| Steven Vinckier | Accepted |
| Tanya Vladimirova | Accepted |
| Fischerová Mária | Accepted |
| Petra.Urbanova@mzp.cz | Accepted |
| riccardo.boschetto@isprambiente.it | Accepted |
| SCRIBE Chrystel | Accepted |
| Emmanuelle Gratia | Accepted |
| PAVOINE Alexandre | Accepted |
| Karppinen Tiina | Accepted |
| Levent ALPAR | Accepted |
| Jean-Paul Lickes | Accepted |
| Vaidotas Vaišis | Accepted |
| Stakvilevičiūtė Dalia | Accepted |
| Masalskis Kęstutis | Accepted |
| Ērika Lagzdiņa | Accepted |
| Natālija Cudečka-Puriņa | Accepted |
| Lina Dagiliene | Accepted |
| Ilze Doniņa | Accepted |
| Astrida Miceikienė | Accepted |
| Erika Tauraitė-Kavai | Accepted |
| Zigmas Medingis | Accepted |
| Visvaldas Varzinskas | Accepted |
| Giedrė Ramanauskienė | Accepted |
| Guštafíkova | Accepted |
| Willems, Mandy | Accepted |
| Anvarifar, Flora | Accepted |
| carmela.cascone@isprambiente.it | Accepted |
| Kristin Pille | Accepted |
| Engeľ Marek | Accepted |
| Ida.Bjorkefall@Naturvardsverket.se | Accepted |
| Louise.Sorme@scb.se | Accepted |
| Ines Areosa | Accepted |
| Christina.Jonsson@naturvardsverket.se | Accepted |
| Žilvinas Danys | Accepted |
| m.grodzinska@gmail.com | Accepted |
| Aistė Litvinaitė-Jablonskienė | Accepted |
| Emanuele Mancosu | Accepted |
| Rebecca.Uggla@Naturvardsverket.se | Accepted |
| guia.agostini@isprambiente.it | Accepted |
| Roger Milego Agrás | Accepted |
| Veselá Magdaléna | Accepted |
| carlo.piscitello@isprambiente.it | Accepted |
| Maria Manuela dos Santos Proença | Accepted |
| Mafalda Mota | Tentative |
| Dráb Ján | Tentative |
| Jurgita Užkurnienė | Tentative |
| MANFREDI Simone | Tentative |
| Daniel Montalvo | Tentative |
| Malene Bruun | Tentative |
| Jekaterina Iljina | Tentative |
| Federico Antognazza | Tentative |
| Şeyma UÇAR SEÇGEL | Tentative |
| MASSON Josiane | Tentative |
| Cecilia.Stafsing@naturvardsverket.se | Declined |
| Valya Zhelyazkova | Declined |
| Odile Le Bolloch | Declined |
| Paul Rasque | Declined |
| Dainius Čergelis | Declined |
| Bert Jansen | Declined |
| Ole Torbjørn Nyvoll | Declined |
| Marguy Kohnen | Declined |
| Mary Frances Rochford | Declined |
| HSR2 | Unknown |
| Martin Adams | Unknown |
| PORRA Federico | Unknown |
| BACIGALUPI Barbara | Unknown |
| m.grodzinska-jurczak@uj.edu.pl | Unknown |
| Littkopf Andreas | Unknown |
| Baskutienė Jolanta | present |

**Agenda**

|  |  |
| --- | --- |
| 10.00 – 10.05 | Welcome and context. *By Peder Jensen – EEA* |
| 10.05 – 10.20 | New emerging data streams and their relevance for CE monitoring​  *Presentation by Yoko Dams – ETC /WMGE* |
| 10.20 – 10.30 | Checking-in: How and when do you work with new data streams in your countries?  *Short warm-up with participants.* |
| 10.30 – 11.25 | Presentation of four pilots by ETC/WMGE Team exploring the use of novel data types for CE monitoring purposes:   * Browser fingerprints * Car sharing system * Web scraping electronic and electrical appliances * Repair case study |
| 11.25 – 11.50 | Sharing insights in the use of the new data streams in EU countries.  “Digitalised monitoring of re-use.”  *Presentation by Nina Lander Svendsen -PlanMiljø*.  *Open knowledge sharing with participants.* |
| 11.50 – 12.00 | Closing and next steps. *By Peder Jensen – EEA* |

**Meeting minutes**

After a welcome and setting-the-scene by Peder Jensen and providing more clarification with regard to the goal and the scope of the ETC study, the meeting participants were invited to let the EEA/ETC know about how the use of novel data type outside the official Eurostat framework has been applied or experimented within their country. An overview of the feedback provided:

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Contact** | **Yes/No** | **Comment** |
| **NL** | Kishna, Maikel | Yes | We are working with different novel data streams such as webscraping of firms, data from ebay like platforms, data on policy actions and instruments, and more. |
| **FIN** | Karppinen Tiina | Yes | In the Finnish Environment Institute SYKE we have experimented with new data streams trying to capture the "inner circles" of circular economy. We have particularly started to look into sharing economy and socio-economic indicators. |
| **A** | Karigl Brigitte;  (Baskutienė Jolanta) | No | We are currently working with EU Monitoring Framework for CE. We are not looking into new data streams. But we are very interested to learn from other experiences. |
| **FR** | Chrystel SCRIBE | No | Sorry, I don't have anything to share. In France we have remained very classical in the use of our date sources. |
| **SE** | Sörme Louise | Yes | Statistics Sweden has a project financed by Swedish innovation agency where we look at different data sources, not only at the national level. I am happy also that you find the repair data from Sweden. |
| **IT** | Carlo Piscitello | No | In Italy we are working with classical Eurostat data. |
| **SRB** | Maja Krunic Lazic | No | In Serbia we are working with classical Eurostat data. |

Four case studies have been presented by the researchers. For each case study the objective, approach used and results were shown. The researcher elaborated also on the obstacles they had to deal with and provided suggestions for further research in case the data stream would be used for monitoring purposes.

A few questions raised by the participants and their answers:

1. Browser fingerprints

|  |  |
| --- | --- |
| **Question** | **Answer** |
| Is it possible to identify the country where the device is used? | You as a user can visit the website, and there is no link with the location. We need to look into other information whether or not captured within the same browser fingerprint (such as IP address) that could give an indication of the geo location of the visitor. For this case study, we have not been investigating in this. For now, we only know that the website provider is based in Montreal, Canada. |
| As I understand there is no information if a customer has intention to change the mobile device in a near future and what happens with the old device? This is " invisible" collection of data? Thank you. | Exactly, maybe a web scraping on used devices sold could be conducted in order to find indications on those aspects. |
| Could mobile phone providers provide such fingerprint data? | Yes, but a collaboration with these companies will be needed, and hence, you would need good argumentations for that. Because they won’t share the data with you without any reason. |
| How well are we covering the market for mobile devices, if we would like to produce an indicator highlighting shifts in the market. Now we lack Apple devices, but maybe there is no need for tracking those, as the android devices give already a quite fair insight in the market trend? | From a technical perspective, we could redo the analysis, but focusing on the identification of IoS devices. The major weakness lies in the set of browser fingerprints as found on the Fingerbank website which is not representative. Another issue that might pop up (but it have not been investigated so far) lies in the legal aspects. |
| How can the data be used for reporting? | The data cannot be used for that yet, but there are possibilities with respect to location detection. A fingerprint contains much more information than the parts we have worked with for this study. |

1. Car sharing – case study for Germany

|  |  |
| --- | --- |
| **Question** | **Answer** |
| Do you observe a pattern of increasing general user number in a car sharing platform over the last couple of years? | For Germany, we observed a growing trend in the numbers (except for 2020, not sure yet what the explanation is here). The grow is mainly driven by the new systems which are station independent. |
| Considering that the cumulative interest on carsharing a function that is always increasing, won't that eventually hide decreases in use of carsharing, if used as a proxy? | If we want to develop an indicator out of this, we would need to look for the right variable. We decided to show the cumulative numbers as from those a relationship between Google Trends and users of car sharing systems is easily derived. For developing a proper indicator it will be more opportune to look at change rates of search interest and how changes in search interests relate to the numbers we have from statistics. Alternatively, machine learning algorithms could be used to develop an estimation for the true numbers out of the number we get from Google. |
| Is it possible to use these data (with additional research) to get a feel for the extent to which car sharing is a substitute for buying/owning cars? | The answer is no. We will have to use additional data here. From the data analysed, we can only derive information on how strong the interest is in car sharing and on registrations and how they are involving over time. We would need to compare this with car ownership data and/or data from surveys to look into this interesting relationship. But it will anyhow be difficult to identify shifts in consumption patterns. |
| The number of users of the car sharing platform, can these number of platform use depend on the export of used cars? | It would be interesting to look at these different trends and compare them. But we cannot conclude this out of the data streams analysed. |
| How can the analysis of Google Trends data be applied to other member states and/or other topics? | In Germany we have the from the branch association. But this is not the case for other countries. Here, Google Trends data could be used to give a grasp of what is going on with respect to car sharing. With Google trends you can zoom in (on regions within countries) and out (compare different regions or countries. And it is a time series, so can check what effects a certain campaign that has been running over a specific period have had. In conclusion, the Google Trends data is much richer than the yearly data provided by the branch association. |
| How did you collect and analyse the data technically? | Google trends data: the most time-consuming step was finding the right keywords. (People won’t use “car sharing” in their search, but rather provider names for example.)  The data were (pre-) processed and analysed using Python. |

And feedback

* Karppinen Tiina: In Finland we have looked into similar data available on city bikes, but this is still in the research phase.
* Willems, Mandy: In the Netherlands we monitor carsharing. Please see [www.crow.nl/dashboard-autodelen/](https://www.crow.nl/dashboard-autodelen/) Dashboard Autodelen - CROW
* De kennispartner voor (decentrale) overheden, aannemers en adviesbureaus. Bij CROW staan samenwerken, groei in kennis en het delen van kennis centraal.
* Mičuda Ján: I think linking the subscription data with car purchase data is tricky. Correlation does not necessarily mean causation. A decrease in car sales might mean that there is a disruption in supply chains, limited production and scarcity on the market (like right now). It does not necessarily mean that people make the conscious decision of switching from owning a car to using a car-sharing service. Although it is very tempting to draw such a conclusion.

1. Webscraping EEA

|  |  |
| --- | --- |
| **Question** | **Answer** |
| Have you managed to break down the Youtube data per product group? | Yes, you could do that. The examples shown concentrated on phone repair. But it depends on the query you use. You can zoom in a devices, brands or even components (for example “screen broken”). We have to select those queries or search terms that are relevant in light of the study conducted. |
| General question: has the EEA explored the collaboration with Google to obtain further access or premium usage of APIs, for instance? | No. So far, the EEA is exploring the usability of novel data types for monitoring purposes. From 2022 on, there will be a new ETC on data and data systems. The idea is to bring in much more expertise in these fields, and put it available for the other ETCs. In this constellation, we will need to come with what could be a business plan first. After that the ETC on data could help us setting up automated information streams, if we consider them relevant. |
| EEE producers will be required to make repair information accessible, is it possible to collect the number of downloads of repair information from producers webpages? | If these producers website would become more and more available, Youtube will be less relevant. This is definitely one of the pitfalls of this approach. We have to continuously follow up on changed environments. |
| If one or more indicators were defined based on Youtube data, how far back into the past can they be computed? | From the web scraping you receive a snapshot of parameters such as number of views, number of likes for that specific moment in time. For building a time series you will need to redo the scraping on a regular basis. |

Feedback:

* Karppinen Tiina: So interesting presentations and examples! It seems to me that we would definitely need both public-private collaboration in order to realise this kind of monitoring as well as surveys to general public to analyse the actual trends from the sustainability point of view.
* Małgorzata Bednarek: More video views may mean that this appliance breaks down and needs to be repaired more often – producing an appliance that is easily or more often to be broken does not seem to be in line with the CE but it may draw the interest in the videos.

1. Repair data

|  |  |
| --- | --- |
| **Question** | **Answer** |
| As I remember the Swedish E-governmental data source it was divided in to different categories of products. Correct? Could be interesting to show! | Yes, they distinguish between two different categories (RUT and ROT as they name it). More details can be provided here, if interesting. |
| How can public authorities support better logging of repairs? | Policy makers could stimulate the repair cafés organized and the communities organising them. But they could also come up with a data standard commonly agreed upon, and hence further aligning and improving the data produced. Thirdly, policy makers could push for having these data on repairs added to the “product passport” as currently under development by the EC. |

Currently, PlanMiljoe is conducting a study for the Nordic Region on how waste prevention could be monitored digitally. On the one hand the researchers are looking for novel data types that could be harvested. On the other hand, data streams that could be unlocked by using a digital technology, is also considered. Researcher Nina Lander Svendson showed how reuse could be measured by looking at data gathered from online second hand platforms. By asking the sellers of used products whether the product is sold (or simply removed from the platform) and combining it with the average mass of the corresponding product category, a translation towards avoided environmental impacts could be achieved. Although cautiousness is needs since strong assumptions are needed and secondary effect such as rebounds might occur.

|  |  |
| --- | --- |
| **Question** | **Answer** |
| How can you make sure to capture the successful ads before they are taken out by the sellers (or by the platform)? | Some platforms ask the seller whether the product is sold are not. This way, you can be sure that the product is simply removed from the platform (and offered on another one). But this is not a common practice across different online platforms. |
| Are the online platforms also collecting the residual values i.e. prices paid for the used products? | No, but potentially it is possible to combine it with economic values using Machine learning for example. Putting a correct price on the used products is really important. Both, prices that are very low or very high could create suspicion among potential customers. Additionally, price setting is tricky, as it is very hard to estimate the residual value of these products. (There is no framework.) |
| People can also give products away for free. In this case, no “official” transfer is registered somewhere. How do this disturb our estimations of the magnitude of reuse in a country? | Sometimes Facebook is used for finding somebody willing to accept the goods. But Facebook is very reluctant sharing these data. But you still can scrape Facebook activities for example. To understand the full range of second hand, we need to continuously monitor the market and try to look for information streams for each option whether commonly used or newly created. |

The meeting was closed by Peder Jensen, highlighting the development of the EEA experimental dashboard (for which a follow up meeting will be held soon).



European Topic Centre on Waste and Materials   
in a Green Economy

Boeretang 200

BE-2400 Mol

Tel.: +14 33 59 83

Web: wmge.eionet.europa.eu

Email: [etcwmge@vito.be](mailto:etcwmge@vito.be)

The European Topic Centre on Waste and Materials in a Green Economy (ETC/WMGE) is a consortium of European institutes under contract of the European Environment Agency.

1. ShareNow is a joint venture of Daimler AG and BMW, and was formed in 2019 by merging the former competitors car2go and DriveNow (Wikipedia, 2021b). [↑](#footnote-ref-2)
2. Miles and Sixt share are listed together on position three of the biggest car sharing providers in the source provided. However, no direct connection between these two providers could be found on an online research. [↑](#footnote-ref-3)
3. The company was rebranded in 2006 in the course of restructuring (Wikipedia, 2021a). [↑](#footnote-ref-4)
4. A car sharing brand by Deutsche Bahn Connect. [↑](#footnote-ref-5)
5. Not available, see text. [↑](#footnote-ref-6)
6. Availability of the vehicle for bookings, regardless of the fact whether the vehicle is actually booked or not. [↑](#footnote-ref-7)
7. https://en.wikipedia.org/wiki/IFixit [↑](#footnote-ref-8)
8. European Commission. (2019). *The new ecodesign measures explained.* Brussels.

   <https://ec.europa.eu/commission/presscorner/detail/en/qanda_19_5889> [↑](#footnote-ref-9)
9. downloadable from: <https://www7.skatteverket.se/portal/apier-och-oppna-data/utvecklarportalen/oppetdata/Rot-%20och%20rutbetalningar,%20statistik> [↑](#footnote-ref-10)
10. <https://standard.openrepair.org/> [↑](#footnote-ref-11)
11. Considered as the aggregation of subsectors:

    Renting and leasing of construction and civil engineering machinery and equipment;

    Services to buildings and landscape activities;

    Repair of personal and household goods [↑](#footnote-ref-12)
12. A country level CE indicator could compare national statistics (number of repair events organized, average success rate) against an European average.; If indicators are drafted on a value chain level, the overall success rate might give an indication of the importance of repair within the value chain. [↑](#footnote-ref-13)