Examining Web User Flows and Behaviours in CLARIN Ecosystem

Go Sugimoto
ACDH-ÖAW
Vienna, Austria
Go.Sugimoto@oeaw.ac.at

Abstract

This article attempts to draw a map of the user flows and behaviours in the multi-layered CLARIN's web structure by cross-examining the dynamic movements of different types of users within (and outside of) the CLARIN domain. In particular, the user traffic of several websites is analysed including the main website, various CLARIN web applications, and the partner websites, as well as the use of single sign-on. Consequently, this project is able to uncover the user interactions in the context of the large web ecosystem rather than those of an each individual website. The evolution of the web traffic over a year reveals a comprehensive overview of the characteristics of the end-users and provides a clue for the next strategic decisions over the CLARIN's user-oriented services and business sustainability. This preliminary research also proves the potential of web analytics for Business Intelligence for measuring the impact of the aggregation services and research infrastructures in cultural heritage and digital humanities.

1 Background – the CLARIN ecosystem

One of the strategies of CLARIN is to create and maintain an infrastructure which is financially, technically and organisationally sustainable in the long-term[1]. It is, therefore, essential to collect and analyse data about its performance and implement objective evaluation which would determine the course of its sustainability. In particular, as CLARIN's core activities are technically-oriented, offering a number of web-based services to the research community, critical evaluation of end users is necessary to check its performance in the long term and to make sensible decisions for the operation of CLARIN. This area of research is generally called Business Intelligence (BI). According to Chugh and Grandhi (2013), the BI is the process of applying tools and techniques to gather and analyse data from multiple sources, to create knowledge that helps in decision-making.

Several evaluations have been conducted for CLARIN in this respect. For example, Eckart et al. (2015) examined the statistics of the Virtual Language Observatory (VLO)[2], attempting to explain the user behaviours. Being a part of their technical development of the VLO, this analysis concentrated on the impact of the change of its design and functionality. Two survey periods were defined to examine the consequence of the interface improvement which took place between the survey periods. Subsequently they observed interesting phenomena relating to the user requests especially on full-text and facets searches. Sugimoto (2017) instead provides more comprehensive research on this topic. He conducted a detailed analysis of web traffic on the VLO from 2014 to 2016, taking into account the number of visitors, visit duration, and frequency, to search keywords, social networks, and downloads, as well as segmenting different user groups such as country. It covers most of the default Piwik[3] analysis views. Although there are challenges to dealing openly with sensitive information about the performance of a

[1] https://www.clarin.eu/content/mission-and-strategy (Accessed on 2018-03-19)
[2] https://vlo.clarin.eu/ (Accessed on 2018-03-19)
[3] https://piwik.org/ (Accessed on 2018-03-19)

community website, it unlocked a potential (or need) of Open Evaluation for publicly-funded research infrastructure.

However, there are two major aspects lacking in those contributions. First of all, they are limited to a single website. Secondly, they focus on a frequently-discussed technical service of the research infrastructure. With regards to the former, we should remember that CLARIN offers many services and websites for different purposes. Therefore, it is not sufficient to study a single website in order to evaluate the technical infrastructure as a whole. Indeed, a similar approach was taken by Culture24 in the UK (Finnis et al. 2012). They recognised that cultural heritage websites and their visitors can be more adequately assessed and understood by knowing the web use within the entire sector. Thus, major museums and cultural institutions including the British Museum, the National Gallery, Tate, and Kew agreed to share some basic statistics of their websites. The interesting initiative made it possible to standardise the datasets of web access across British heritage institutions and analyse the landscape of their web users. It may have been the first time that an overview of the web traffic within a larger sector was revealed, which massively contributed to the understanding of the bigger picture of emerging museum and heritage business on the Internet. As for the latter, CLARIN's success indicator should not be determined only by technical web applications, but by many other social and organisational services around. In particular, the main website of the infrastructure (CLARIN ERIC: European Research Infrastructure Consortium) should be included.

For those reasons, this paper (re-)evaluates the CLARIN services from a different angle. It takes a holistic approach to capture the traffic of end-users across various websites and applications as well as national centre websites in an attempt to better understand more global aspects of the "customers" of CLARIN. To this end, let us first analyse the CLARIN's web environment.

Although the individual websites of CLARIN are relatively simple, the whole web structure is multi-layered with regard to user movements (Figure 1). The most obvious website is clarin.eu. It is often an entry point for the existing and new users, mainly serving as a communication and dissemination website. It does not only offer the basic information (the missions, people, participating institutions etc.) and updates news and events, but also links to useful websites and services inside and outside the CLARIN. In addition to the main website, there are many web applications developed by the CLARIN developers such as, the VLO, Content Search Aggregator[4], and WebLicht.[5] They are useful research tools and are deployed either in the subdomains of CLARIN or its partners domains, therefore, truly making CLARIN a distributed infrastructure. The users jump from the main website, or directly go to, those services to start their research. Although more limited, the users also navigate between the CLARIN services and the partner websites. Many CLARIN national consortia have websites dedicated to providing domestic information, including CLARIN DK[6] and LT[7]. Moreover, CLARIN centres may have their own websites often placing the CLARIN logo to suggest their connection, for example, the Center for Sprogteknologi[8] in Denmark and the CLARIN Text Laboratory Centre[9] in Norway.

As such, there are at least three major entry points to the CLARIN websites (the main website[10], the CLARIN applications, and the national consortia/centres) and the movements of the users among those websites are complex. The author gives the name, "CLARIN ecosystem", to refer to the full picture of those websites within the CLARIN community. In the sense that we analyse the web access and user flows within the CLARIN community, our approach is different from Culture24, which focuses on completely independent museum websites.

Among the web applications, VLO is probably the most typical case of the CLARIN ecosystem. Therefore, it deserves the name of "VLO ecosystem" on its own. It is a resource discovery portal service to search and locate the linguistic data and tools that the CLARIN consortium members hold, hence it merely collects metadata as an aggregation service provider. Van Uytvanck et al. (2010) describes that

[4] https://spraakbanken.gu.se/ws/fcs/2.0/aggregator/ (Accessed on 2018-03-19)
[5] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page (Accessed on 2018-03-19)
[6] http://info.clarin.dk (Accessed on 2018-03-19)
[7] http://clarin-lt.lt (Accessed on 2018-03-19)
[8] http://cst.ku.dk/ (Accessed on 2018-03-19)
[9] http://tekstlab.uio.no/clarino/ (Accessed on 2018-03-19)
[10] It should be noted that there has been no detailed research on the web statistics of the main website, except some general facts and numbers demonstrated, for example, in CLARIN Annual Conferences as well as usability studies.

it tries to give a consistent online overview of the data that is available at a variety of computing centres. Using VLO, the users are directed to the repository of a data provider where the resources they find in the VLO search engine are stored.

Alongside such user streams, the CLARIN's single sign-on services will be examined in order to check the user behaviours by different types of the users including anonymous, the CLARIN registered, and academic users. The objective of this paper is, therefore, to unveil the interactions of various types of users in the large ecosystem which could not be recognised by the previous research based on the observation of a single website.

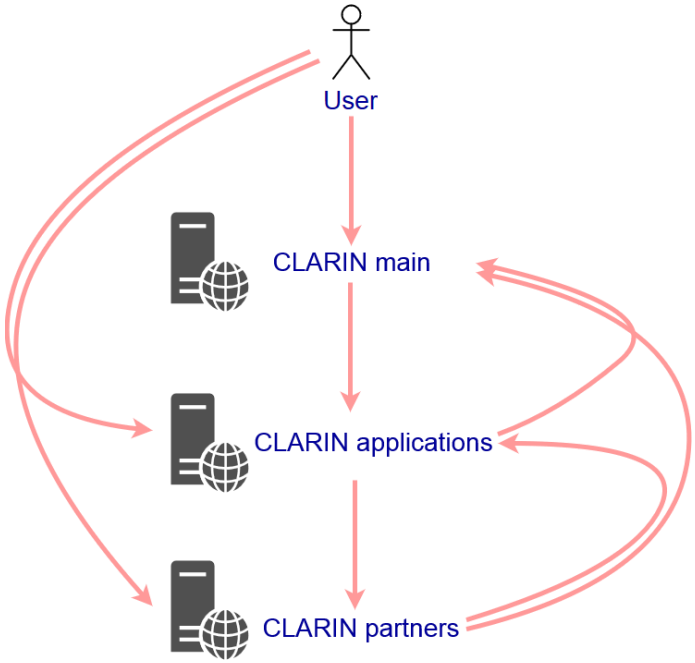

Figure 1. Multi-layered CLARIN web structure ("ecosystem") and user access

2 Methodologies

The data range of this project is between February 1st 2016 and January 31st 2017, taking into account technical limitations and comparative studies (Figure 2). While Google Analytics is also used to record the traffic of the CLARIN main website, inevitably, Piwik was our choice to analyse the data, as it is the only GUI tool which keeps tracks of all the CLARIN websites that concern us. However, Piwik has been collecting the statistics of different websites since varying points in time. As the main website only started to use Piwik in 2016, we set the beginning of its recording more or less as the beginning of our analysis period. The first half of the period corresponds to the last quarter of the survey by Sugimoto (2017), which might be also useful, if the need of cross-analysis emerges in the future.

In order to reconcile the broad spectrum of the CLARIN's web structure, the author inspects the following websites: the main web-site, VLO, WebLicht, the Content Search Aggregation, the Discovery Service, and the Identity Provider. Although this does not include all of the CLARIN websites, it is assumed that it covers most critical ones, representing what CLARIN offers on the web.

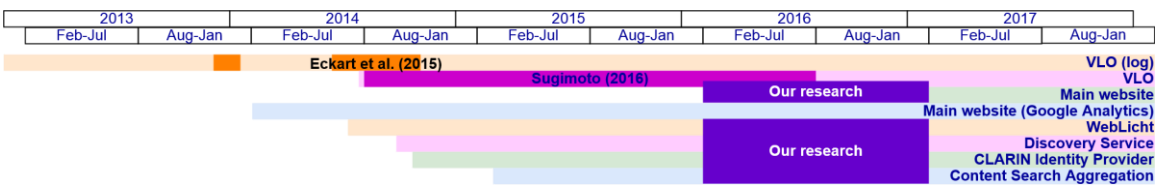

Figure 2. Research period coverage (Available data periods are represented in light colours and research periods in dark colours)

The transition view of Piwik is primarily used to analyse the user flow, in combination with other statistical data (Figure 3). It is a powerful tool, able to visualise from where users come to a certain web page and to where they move. As a single website comprises many webpages, there would be potentially

hundreds of views to check the user movements in this way. To avoid interpreting such a large amount of data, it is decided to select some hubs or junction points of user flows. The transition view also allows us to distinguish internal flows (inside the domain) from external flows (outside the domain), thus is very useful to understand the user behaviours. In addition, the classifications of traffic enables us to divide users into separate groups such as the users visiting by search engine or direct entry, as well as the users who downloaded a file or quit the web page. Such footprints of users would provide interesting information for improving CLARIN services.

3 Analysis on the CLARIN (especially VLO) ecosystem

3.1 At the main website -entry gate to CLARIN/VLO ecosystem

First of all, the entry points of the CLARIN ecosystem are examined. Figure 3 illustrates the user flows of the main website at its home page (i.e. clarin.eu). 21,945 page views are recorded in the period, in which 23% are from internal webpages, 18% from search engines, 10% from web referrers, and 36% from direct entries. Within the search engine flow, keywords like "clarin", "clarin eric", "clarin eu" and "https://www.clarin.eu" are extremely prominent with 85.6% in total. This implies that most users already knew CLARIN by name, or even the URL, and did not find it by coincidence, for example, when searching for linguistic information. As for the outbound paths, 51% of the users remain on the main website, of which 12% are through to Services, 11% to Events, 8.8% to Participating Consortia, 5.5% to Clarin-in-a-nutshell, and 5% to Users. In addition, 2.8% visited another website, whereas 40% exited (i.e. no more actions by the user). The statistics proved the importance of the VLO as one of the CLARIN's primary services, as it gained 30% of the Outlinks of the visitors. The CLARIN Germany (3.8% for clarin-d.de) seems to be successful in attracting users from other countries.

It is possible to try to estimate the existing users discussed above more in detail. Firstly, the external access to the website should be the amount subtracting reload and internal pages (21945-5089-2039=14,817). The total amount of possibly existing users would be the sum of the search engine access with keywords related to CLARIN and direct access (135+9+5+3+7876=8028). The external access divided by existing users is, therefore, 54.2%. This would be the minimum amount as other channels of access can be observed. This figure can be compared to the more conventional statistics of repeating visits. Piwik recorded 10,000 visits for the same page views as Figure 3, when filtering visits more than once, which is 45.6% of the amount without filtering (21,945). It is not easy to explain the gap. Although Sugimoto (2017) interestingly investigated the black box aspect of Piwik (which would be applicable to any other Web Analytics), both the simple methodology of estimation here and the access handling and recording mechanism of such software are the factor of discrepancy and error. Still, this quick experiment seems to be the only way forward to try to understand the nature of Web Analytics and to adequately and systematically evaluate the web traffic.

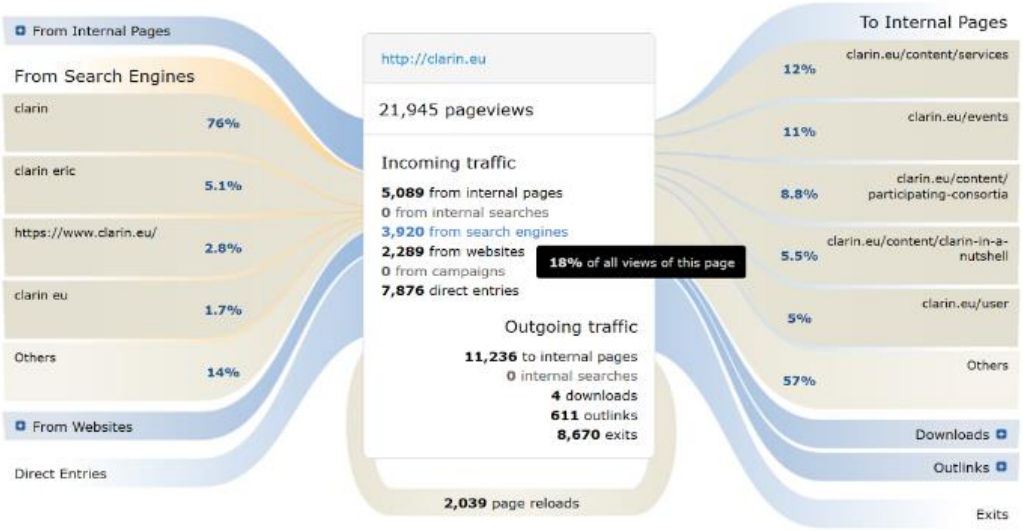

Figure 3. Transition view of the home page of the main website.
(It shows the web page of analytical interest (http://clarin.eu) in the centre, where the user was before on the left, and to where the user went afterwards on the right.)

Another point of interest is the Participating Consortia page.[11] It lists national consortia of European countries and observers. As described in section one, each consortium may have their own website, so that we can check what consortium receives visitors from this page.

According to Figure 4, outlinks are most represented by CLARIN Germany (12%), Austria (10%), Italia (9.6%), the UK (6.7%), and Latvia (6.1%). In contrast, although the total volume of traffic is 4 times less (i.e. 130) than outgoing traffic (i.e. 522), the incoming traffic from external websites originates from other CLARIN consortium domains. They are CLARIN Slovenia (50 and 25%) and Greece (13%) alongside Wikipedia Germany (13%). When we look at big announcements of national consortia joining CLARIN, there are three relevant countries in the survey period: Latvia (1st of June 2016)[12], Hungary (1st of August 2016)[13], and France (1st of February 2017)[14]. Access to the Latvian website may be explained from this data, whereas the reasons for traffic to other popular consortia are unclear at this stage, as is the absence of Hungary and France. As mentioned above, the participating consortia page is one of the most visited web pages from the main page, so that it would be wise to provide useful and informative content about who the members of CLARIN are besides promoting the national websites. With regard to the internal web pages, nearly half of the visitors (48%) comes from the CLARIN home page, which is naturally expected.

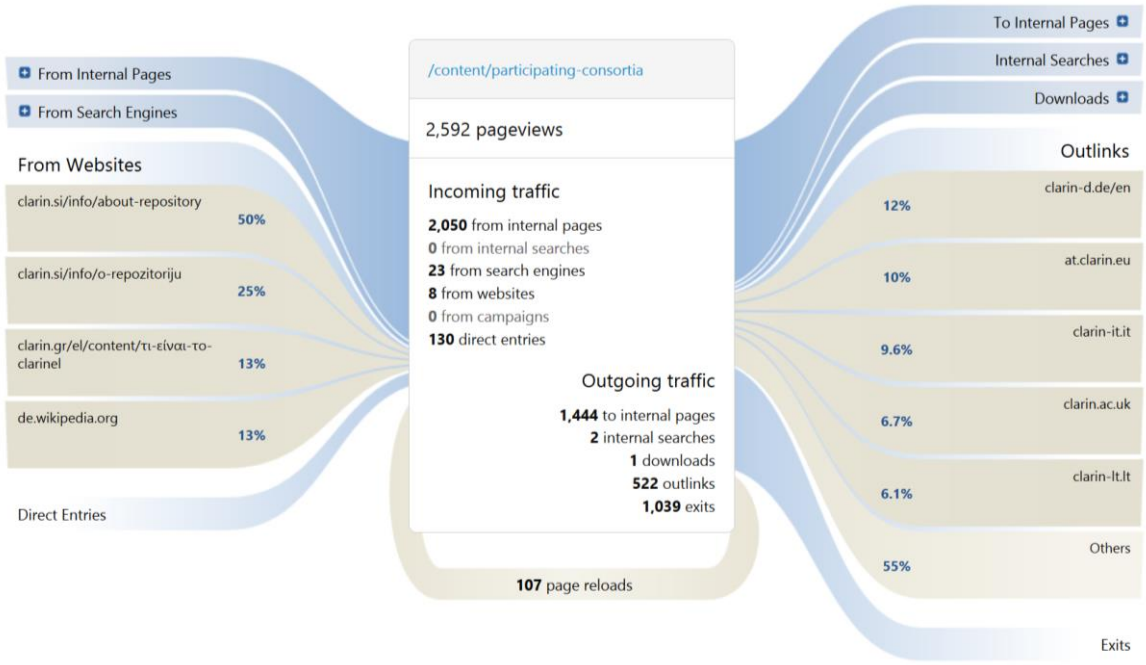

Figure 4. Participating consortia page user flow

The services section of the main website is also a decision-making point[15]. This introduction page contains several links to specific services and applications, therefore, allows us to identify the trend of user interests in those services. Figure 5 shows an extremely high percentage of inbound and outbound traffic for the internal website – 92% (3533) and 89% (3402) respectively. Also, given the introductory nature of the content, the data confirm that it is a typical walk-through page of the main website. The decision-making of which links to follow comes into play in our analysis. Within the outgoing flow, the main web page (clarin.eu/portal and clarin.eu) is prominent, but the VLO (9.8%) and the Language Resource Inventory (7%) are also visible among the top 5. Whilst the former is anticipated (see also below), the latter suggests that the users are interested in the LINDAT service on which the Language Resource Inventory is based. Although the incoming flow from external websites is highly limited, there are interesting facts that a few websites have a direct link to the service section page such as the University of Münster and Academic IT Research Support team of the University of Oxford. This is a case

[11] https://www.clarin.eu/content/participating-consortia (Accessed on 2018-03-19)
[12] https://www.clarin.eu/news/latvia-joins-clarin-eric (Accessed on 2018-03-19)
[13] https://www.clarin.eu/news/hungary-joins-clarin-eric (Accessed on 2018-03-19)
[14] https://www.clarin.eu/news/france-joins-clarin-eric (Accessed on 2018-03-19)
[15] https://www.clarin.eu/content/services (Accessed on 2018-03-19)

of a small fraction of user flow, but Piwik has proven useful to analysing what referrals exist and how the users enter the CLARIN ecosystem.

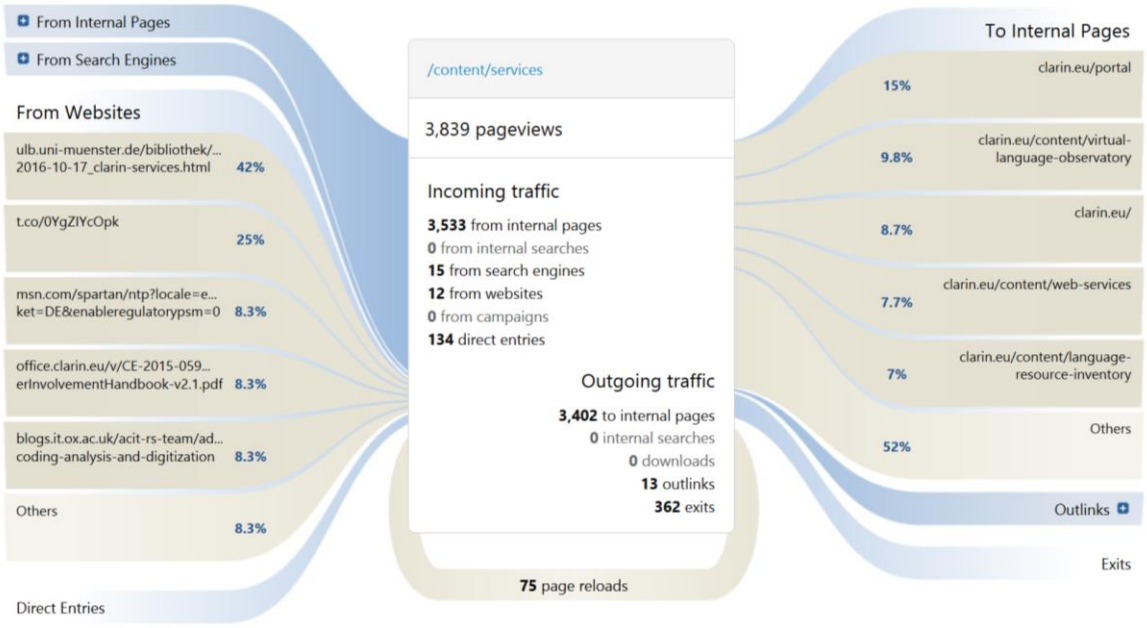

Figure 5. User flow at the service section of the main website

There is a VLO introduction page on the main website which would be one of the main gates to the VLO (Figure 6). 77% of all the visits went to the VLO, so that most users pass this connection point to arrive at the VLO. 44% of the users find the web page from the service section of the website, while the other routes are rather limited (internal search 0.1%, website 4.7%, direct entry 12%). The relatively high number of entries via search engines (21%) suggests that the users know the VLO, because their search keywords include specific terms referring to the VLO or CLARIN. The user flows from the VLO to the CLARIN centres are much more complex and the examination is in progress. Although understandable, it is a pity that we have no access to the statistics of the CLARIN centres. If the access permission is somehow granted, it is possible to examine the complex VLO ecosystem in a similar way that Culture24 was able to do. What we suggest is to share a subset of the whole data in the form of spreadsheets export, instead of the unlimited access to Piwik and/or Google Analytics. Collection of such data dumps from various centres will shed a light on the understanding of the navigation of the CLARIN users.

A part of the problem is that the individual URIs of the centres need to be checked and the use of Persistent Identifiers (i.e. Handle[16]) makes it untraceable without manual clicking and checking of all the URIs recorded. Nevertheless, apart from Handle, the University of Leipzig (2.8% of all Outlinks) and the SIL International (2.1%) received more visitors than others.

[16] https://www.handle.net (Accessed on 2018-03-19)

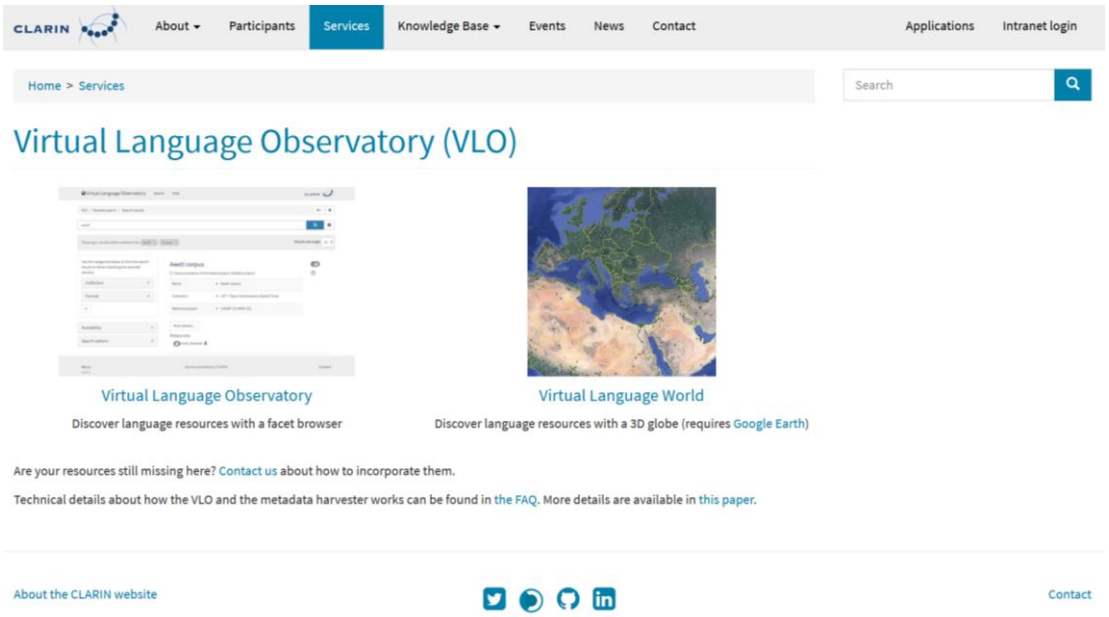

Figure 6. VLO introduction page at the main home page

Before moving on to the VLO itself, let us double-check in what position the VLO is. Figure 7 outlines the top 10 highest number of outgoing access from the entire clarin.eu domain (i.e. not only from the clarin.eu home page, but including it). The VLO tops the ranking as 11.2%. Other VLO related technical services such as centres.clarin.eu (Centre Registry 4.9%) and catalog.clarin.eu (Changing address including Component Registry etc. 4.5%) follows Handle persistent identifiers (6.9%). It is also notable that LINDAT and WebLicht (see section) are among the pages visited frequently by the users.

	URL	Unique Clicks	Percentage
1	vlo.clarin.eu	998	11.2%
2	hdl.handle.net	615	6.9%
3	centres.clarin.eu	439	4.9%
4	catalog.clarin.eu	404	4.5%
5	www.clarin.eu	386	4.3%
6	infra.clarin.eu	255	2.9%
7	lindat.mff.cuni.cz	227	2.6%
8	www.clarin-d.de	215	2.4%
9	docs.google.com	181	2.0%
10	weblicht.sfs.uni-tuebingen.de	118	1.3%

Figure 7. Top 10 outlinks from the whole clarin.eu domain

3.2 At the VLO

From the VLO's point of view, the trend of in- and out- channels is different (Figure 8). 22% of the visits originate from web referrers. 600 out of 1102 visits from websites (54%) are the VLO introduction page (with additional 5.8%). Interestingly, the Stackexchange website has a post about a Korean language corpus and the VLO is mentioned. As a result it gained a high rate of access (7.4%) during this period. Similarly, 5.0+ % are observed due to the University of Vienna offering a Moodle link to the VLO. Unlike the main website, a low number of users landed with the VLO via search engines (4.8%). 29% of the users find the website directly. Regarding the outward traffic, we can see a clear trend for *Korean* probably caused by the abovementioned stream ("korean" (1.7%) and "korean corpus" (2.6%)). At the first glance 31% of the users who went through to internal pages may have done so by browsing, because the VLO is a search engine which, in principle, should increase internal searches (24%). However, this assumption cannot answer why internal searches are less than page browsing. When it is discovered that the internal pages contains the URL syntax pattern such as "vlo.larin.eu/search?1", the

classification by Piwik becomes slightly dubious. It is nevertheless important to note that unlike on the CLARIN main website, much higher numbers were recorded for internal searches both in and out directions of the traffic. Yet another puzzle piece is the difference between 637 (inbound internal searches) and 1218 (outbound internal searches) as well as the existence of both identical search keywords and different ones. In general, more iteration of observations, analyses, and experiments would be needed to solve this kind of mystery, for example, by understanding the details of the mechanism of auto-generated URLs in the VLO, as well as what the Web Analytics records and classifies. In the meanwhile, 21% exited without doing anything.

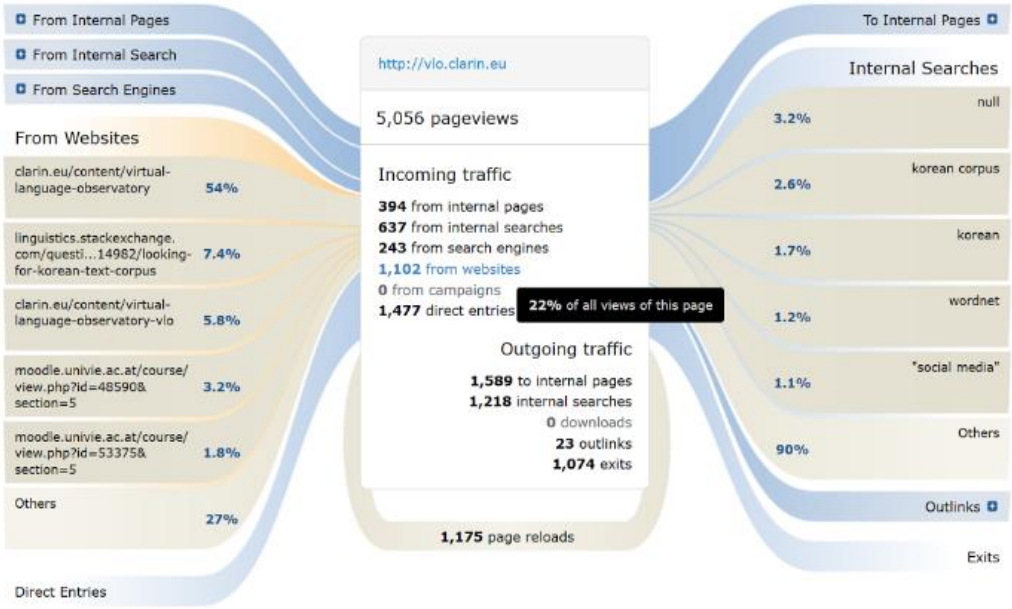

Figure 8. User flow at the home page of VLO

On the other hand, overall search keywords left some clues on the user needs (Figure 9). Interesting cases of Korean have already been introduced. In addition, some users search something very specific such as "hzsk" (0.9%, probably intended for the CLARIN B centre of Das Hamburger Zentrum für Sprachkorpora (HZSK)[17]), "GECO" or "geco" (0.9%, also intended for IMS GECO Datenbank provided by the CLARIN B centre of Universität Stuttgart[18]), and "germanet" (0.3%, also intended for the service by University of Tübingen[19]). It is obvious that they look for German data and tools. It seems that such access is made by the CLARIN's internal users, rather than the experts outside CLARIN who know exactly what CLARIN offers. The tendency towards language names cannot be ignored and this trend was also found during the two years of Sugimoto's analysis (2017).

[17] https://corpora.uni-hamburg.de/hzsk/ (Accessed on 2018-03-19)
[18] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.html (Accessed on 2018-03-19)
[19] http://www.sfs.uni-tuebingen.de/GermaNet/ (Accessed on 2018-03-19)

KEYWORD	SEARCHES	SEARCH RESULTS PAGES	% SEARCH EXITS
null	2.1% 86	1.2	12%
hzsk	0.9% 39	3.2	67%
korean corpus	0.9% 36	3.6	61%
corpus	0.6% 23	3.3	48%
russian	0.6% 23	2.3	83%
treebank	0.6% 23	2.5	57%
korean	0.5% 20	2.6	25%
geco	0.5% 19	1.9	53%
german	0.5% 19	3.3	47%
GECO	0.4% 17	1.9	88%
wordnet	0.4% 16	3.1	31%
chn	0.4% 15	1.4	100%
dutch	0.3% 14	4.2	50%
germanet	0.3% 14	2.1	50%

Figure 9. Search keywords used in the VLO

It is also very easy to learn what the users downloaded (Figure 10). However, as the number is significantly lower than the visits in total, it was decided to display only the highest ranking URIs in this paper. It is perhaps fair to mention that the trouble of this type of analysis is that 120 links have to be manually clicked and checked to know exactly what the downloaded contents are about. Although some URIs could give some hints of content in the syntax, opaque URIs, especially persistent identifiers like Handle, make it impossible to guess the content of the target resource. Given that the number of outlinks is much bigger, there are limits for the manual analyses. This is one of the very interesting and unfortunate pitfalls of persistent identifiers in terms of Web Analytics. This paper does not mean to say that opaque URIs should be avoided. Rather, it only suggests that both the creators and implementers of persistent identifiers may need to consider this aspect for improvement or solution in the future, if Web Analytics deploying substantial amount of manual work is considered to be important.

DOWNLOAD URL	UNIQUE DOWNLOADS	DOWNLOADS
vlo.clarin.eu/ - Others	57	59
corpora.uni-leipzig.de/downloads/ukr_newscrawl_2011_1M-text.tar.gz	2	3
vlo.clarin.eu/record?2-1.ILinkListener-cmdi-toggler-link&docId=CLARIN Centres/oai_clarin_pl_eu_11321_270.xml&q=KPWr&index=8&count=13	2	2
vlo.clarin.eu/record?5-1.ILinkListener-tabs-tabs~container-tabs-1-link&docId=CLARIN Centres/oai_clarin_pl_eu_11321_270.xml&q=KPWr&index=8&...	2	2
cocoon.huma-num.fr/exist/crdo/schang/gcf/crdo-GCF_1016.xml	1	2
clarin.phonetik.uni-muenchen.de/BASRepository/Corpora/CH-Jugendsprache/SNF_jspr_I4_S2_001_061019_Beziehungsnetz/SNF_jspr_I4_S2_001_0610...	1	1
clarin.phonetik.uni-muenchen.de/BASRepository/Corpora/SC10/CLARINDocu.zip	1	1
clarin.vdu.lt/xmlui/bitstream/handle/99999/10/ALKSNIS_v2.zip?sequence=1	1	1
corpora.uni-hamburg.de/repository/file:kolas_kolas-1.0-documentation/PDF/andresen-knorr-kolas-dokumentation.pdf	1	1
corpora.uni-hamburg.de/repository/file:kolas_kolas-1.0/ZIP/kolas-1.0.zip	1	1
cts.informatik.uni-leipzig.de/teidumps/pbc/bible/parallel/deu/elberfelder1905.xml	1	1
hdl.handle.net/11041/alipe-000853/ali-baptiste-101227-2.xml	1	1
hdl.handle.net/11041/sldr000758/olac.xml	1	1
sldr.org/logo/LogoOrtolang_small.png	1	1

Figure 10. Top download URIs within the VLO

Those additional (potential) analyses clarify that multi-dimensional analyses, combining user behaviour analysis, in this case, for keyword searching and downloading, with transition analysis, can make a significant contribution to the understanding of the users as the principal creatures of the ecosystem environment.

4 WebLicht and Content Search Aggregator

"WebLicht is an execution environment for automatic annotation of text corpora. Linguistic tools such as tokenizers, part of speech taggers, and parsers are encapsulated as web services, which can be combined by the user into custom processing chains."[20] Consequently, the structure of the website/web application is very different from the main website and the VLO, resulting in no transition view produced by Piwik. In fact, the user flow exists in terms of the data processing chain, but not in terms of web pages. Therefore, we need to look at other statistics. 70% of visits to WebLicht are referrers, while 29% are direct entry. As the CLARIN-D is the developer of the WebLicht, the referrers are mostly from the German domains, except for the top score of "idp.clarin.eu" (29%). Similarly, Germany dominates the visits by country (82%), while there is also interest from Austria (3%), South Korea (3%) and the United States (1%) (Figure 11). WebLicht is perhaps something CLARIN has failed to promote. Although it can handle many languages (for example, there are more than 40 language choices for plain text processing), the service is almost exclusive to Germany, the major CLARIN consortium member. It seems that CLARIN would need to review the outreach strategies of WebLicht in order to go beyond the German niche market. The visit duration is substantially longer (11 minutes 57 seconds on average) than for the VLO (4 minutes 18 seconds) (Sugimoto 2016). 27% spend more than 10 minutes, proving the characteristics of the data processing service. This engagement promises that new users potentially become heavy users.

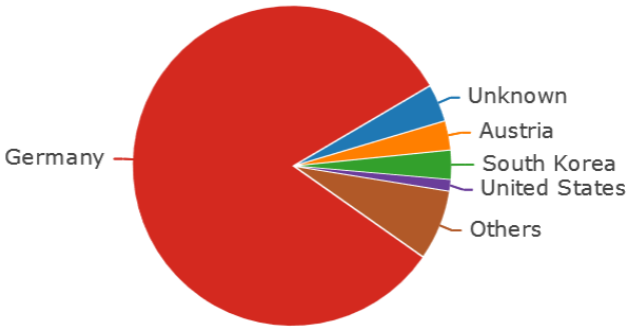

Figure 11. WebLicht visits by country

As for the Content Search Aggregator (Figure 12), the transition view is valid. A high ratio of page reload was detected (39%) in comparison with the main website (9.3%) and the VLO (23%), whereas web referrers come second at 34%. A very low amount or no users arrived internally (i.e. via web pages (0%) and search (2.3%)). In fact, less than 10% accessed from the CLARIN main website. On the other hand, CLARIN-D successfully converted their users to the Content Search users (over 75% of referrers). The implications of those results need to be further investigated. Incoming internal searches indicates the presence of German speaking users, as the most searched keywords are all German including "armut" (poverty, 17.4%), "Forsythie" (Forsythia, a type of shrub, 4.3%), "selbstmord" (suicide, 4.3%), and "diachrone deutsche korpora" (diachronic German corpora, 4.3%). Regarding outgoing internal searchers, there are more varieties, but German words are still the most visible. "Leipzig" (1.6%), "selbsmord" (1.6%), "vom text zur phonologischen aussprache" (from text to phonological pronunciation, 0.8%), "mal eben" (just in a moment, 0.8%) are shown in the highest. The same marketing argument we made for WebLicht applies to the Content Search Aggregator.

[20] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page (Accessed on 2018-03-19)

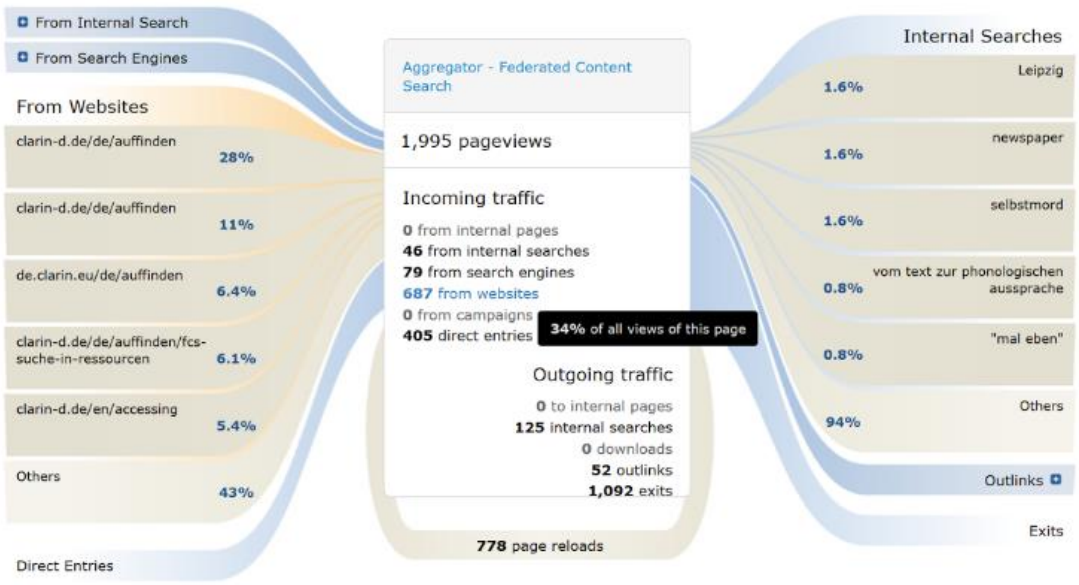

Figure 12. User flow at Content Search Aggregator

Compared to the VLO ecosystem, the situation of other applications is different. It is worth analysing the characteristics of the visits, but the user flow inferred from Piwik is rather limited. We can conclude that although it is necessary to monitor the flow from the main website, those services are rather the end points of the CLARIN ecosystem, thus, it is more productive to analyse the VLO ecosystem in this sense.

5 Identity Providers and Discovery Service

CLARIN provides a pragmatic solution for user authentication and authorisation. The recording of user sign-on and access to web services enables us to explore the statistics of different user types in CLARIN's web space to support our previous analyses. We analysed Identity Provider (i.e. only users with CLARIN credentials) and Discovery Service (i.e. all users trying to access log-in services)

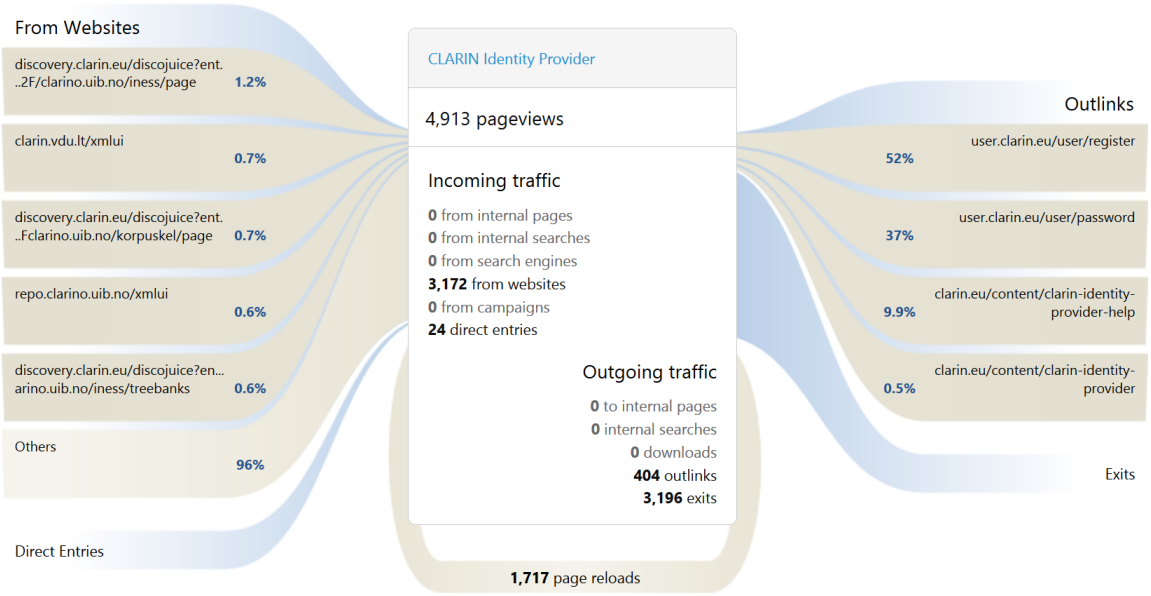

Figure 13. User flow for Identity Provider

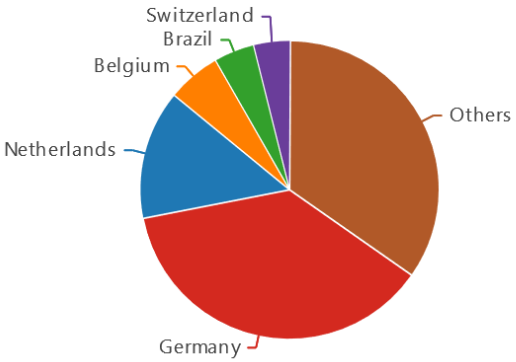

Figure 14. Access by country at Identity Provider

Although the URIs are relatively spread (Figure 13), as for the incoming web referrers of Identity Provider, there are only two countries which gain visibility: Lithuania and Norway. The Lithuanian repository of CLARIN-LT[21] (0.7%) and the Norwegian repository by CLARINO[22] (0.6%) delivered more users to the Identity Provider. The Norwegian boom is further boosted by other URIs of (probably) INESS.[23] In contrast, interestingly those Northern countries do not appear in the access by country (Figure 14). It may mean that German and Dutch users, as well as Belgian, Brazilian, and Swiss users to a lesser extent, are interested to find Lithuanian and Norwegian resources. The assumption was mostly right that among the total amount of 99 views of Norwegian domains within the top 5, the users from Germany accessed Norwegian domains 38 times and users from Norway did 43 times. The large majority comes from those two countries. It is a common phenomenon that users use resources from their own country, therefore, German is something unique in this context, in a way fulfilling the aim of CLARIN to encourage trans-European access. However, it is again true that the German population bias as well as the influence within CLARIN are big. Outbound traffic seems to be rather technical and there is not much from which we can draw conclusions. In case of Discovery Service (Figure 15), it is the Netherlands which dominates the scene for the incoming traffic. Among the URIs, Corpus Hedendaags Nederlands (9.4%+6%)[24], Open Sonar (4.5% and 4.2%)[25], although WebLicht shows strength (9.1%). This is clearly represented in the pie graph depicting access by country (Figure 16). The swap of German and Dutch users is quite dramatic and interesting, but we need more evidence to explain this situation. Again, outlinks contain URIs too technical to mention.

In the meanwhile, both the Discovery Service and Identity Provider have a large proportion of exit (65% and 82% of outbound traffic respectively). This may imply that many users give up access due to this access restriction. In that case the Service Providers may want to reconsider their access policies. While the former acquired 41% from referrers, the latter is at 94%, which is probably naturally high as a sign-on screen appears when a link on a webpage is clicked. It is, however, noted that the technical mechanisms of those services are complicated, making the recording (and interpretation) of the user access in Piwik very tricky. In order to clarify the situation, the next step of investigation would be to carry out an experiment to understand what Piwik actually records behind the user interactions with those CLARIN services, using the Visitor Log function.

[21] https://clarin.vdu.lt/xmlui/ (Accessed on 2018-03-19)
[22] https://repo.clarino.uib.no/xmlui/ (Accessed on 2018-03-19)
[23] http://clarino.uib.no/iness/page (Accessed on 2018-03-19)
[24] http://corpushedendaagsnederlands.inl.nl/ and http://chn.inl.nl/ (Accessed on 2018-03-19)
[25] http://opensonar.inl.nl (Accessed on 2018-03-19)

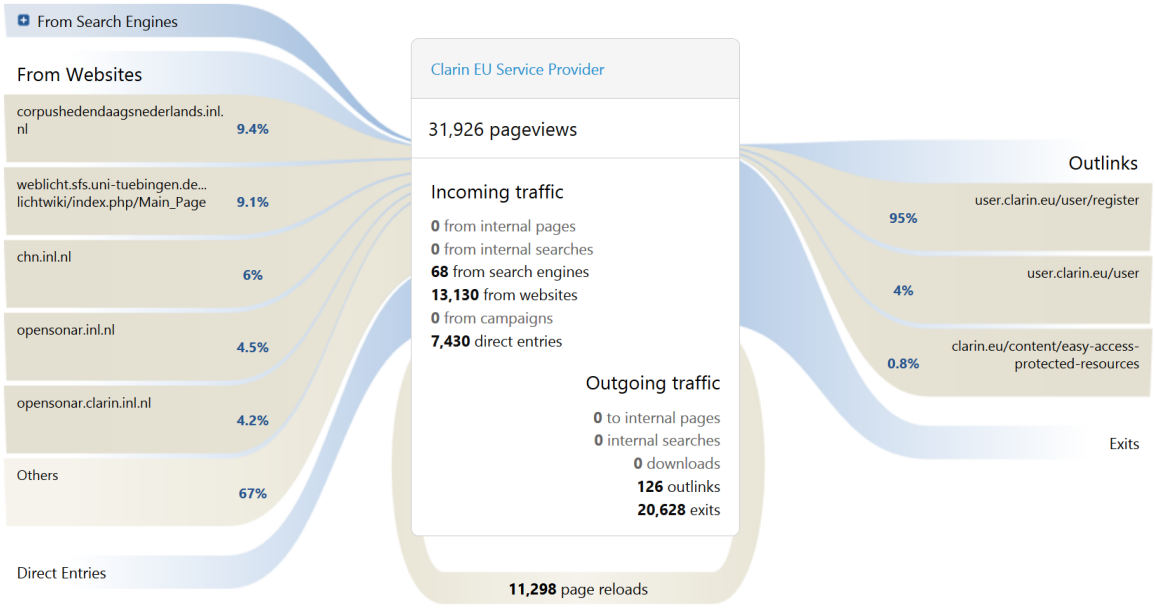

Figure 15. User flow at Discovery Service

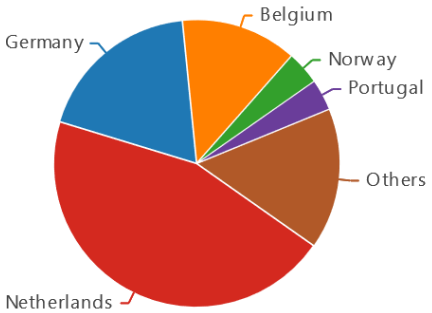

Figure 16. Access by country at Discovery Service

6 Conclusion

The transition view of Piwik in combination with other functions allows us to effectively evaluate the user traffic streams in the multi-layered web structure. It is easy to browse the types of inbound and outbound movements of the users. In particular, an unprecedented amount of statistics about the CLARIN and VLO ecosystem was analysed in detail. Whilst the role of the VLO at the centre of CLARIN's web infrastructure was confirmed, the complex user flow within the ecosystem uncovered some trends of this survey period. Some junction points like participating consortia and the VLO introduction page provide insights into the users moving out of the main website. In general, the existing users, characterised by direct entry and "intended access by search engine" (keywords are used to specify CLARIN and VLO), seem to influence the statistics to a large extent. In addition, the fact that many specific search terms are recorded in the VLO also adds to the evidence of CLARIN's internal community users. It may be still too early to draw a conclusion that CLARIN websites rely on internal community users. However, initial results collected so far are affirmative, even when they are compared to the outcomes of Sugimoto (2017) who suggested a heavy usage of the VLO by a CLARIN partner in Austria[26]. In a way, CLARIN fails to catch attention from external community. CLARIN should definitely consider the transition of heavy user base from CLARIN members to the outsiders. For example, the CLARIN annual conference could be open to a larger community, thus, the infrastructure can be more widely recognised and used. In addition, CLARIN could reduce the internal networking and research mobility between CLARIN centres, and increase workshops and seminars in fringe domains such as

[26] Austria is often regarded as one of the core technical members of CLARIN.

philology and language-related subjects. In particular, the researchers who do not normally deploy computational linguistic methods would need crash courses to obtain practical skills and knowledge to use CLARIN resources and tools.

The high volume of flow from Germany can be seen in different traffic records, but the population bias is not yet take into consideration. Nevertheless, as one of the core members of the CLARIN consortia, Germany hugely influences the web traffic. On one hand, CLARIN benefits from the driving force, on the other hand, the European infrastructure seems to need more effort to expand the user base outside Germany. The value proposition of CLARIN clearly states (CLARIN ERIC 2017) that "as generic infrastructure services can be used across borders, CLARIN members can benefit from the fact that the costs of construction and operation of such services can be shared between members" and "access to CLARIN resources (data, tools and methods) will also lead to more advanced research and open new research avenues across borders and disciplines". For this reason, the reduction of CLARIN activities in Germany and the expansion of CLARIN programmes in less popular countries may be a good option for widening the user diversity. More knowledge transfer from active CLARIN countries to less active countries would also be a new strategy agenda for cross-border synergies. The impact of a sudden increase in particular access paths such as "Korea" became easily visible from the beginning to the end of the access paths, supporting the detailed analysis possibility of Piwik.

Although WebLicht and Content Search Aggregator provided less useful information about the user flow, and are thus not extremely suitable for the analysis of the user movements within the CLARIN ecosystem, they underpin the large amount of German users. Identity Providers and Discovery Services are also particular in the sense that they are the layers to go through to CLARIN services. The analyses revealed that Norway and Lithuania gain popularity, mainly due to the access from Germany. The dramatic swap of the Netherlands and Germany poses a question to be answered.

There are also some areas where further research is needed to clarify the situation and provide correct interpretation. For example, it is a challenge to scrutinise the websites after a major overhaul (for example, Goosen and Eckart (2014) and CLARIN ERIC (2016)). Web addresses may change over time due to the introduction of new underlying software and/or restructuring of the website. Such a change introduces a complicated list of page URLs for transition analysis. It would be wise for the web analytics team and development team to closely communicate about the web development plans, so that the troubles of web traffic evaluation could be minimised and CLARIN's tasks can be more efficiently coordinated, for instance, by extending the members of the VLO Task Force (Haaf et al. 2014). Besides, the marketing strategies created by web analysis and the development of websites could go hand-in-hand for the efficient and continuous improvement of the infrastructure. The tight cooperation would potentially saves money and address the needs of the right users and other stakeholders. Therefore, it is important both in terms of the technical, organisational, business, and financial stability of CLARIN. More cooperate governance[27] needs to be implemented. In addition, the technical mechanism behind authorisation and authentication in relation to the recording of Piwik is still unclear. Moreover, a pitfall of opaque persistent identifiers was recognised. It is not a big problem for the scale of analysis in this paper, but as the web access grows, it would make detailed and interesting analyses more difficult.

The preliminary results of this paper successfully displayed new in-sights into the end-users of CLARIN. In addition, this is probably the first time to synthesise the statistical analyses of both the dissemination website and the web applications of CLARIN in terms of user traffic. Moreover, it is also a reconfirmation that it is important to monitor the statistics over time. A comprehensive implementation of Business Intelligence would require more data from different areas such as financial reports and user engagement reports. Nevertheless, it is hoped that this small research project has brought some ideas about the visitors and environments of the CLARIN's virtual ecosystem in the framework of web analytics and would be a valuable contribution to the development and sustainability of CLARIN.

References

[Chugh and Grandhi 2013] R. Chugh, and S. Grandhi. 2013. Why Business Intelligence? Significance of Business Intelligence Tools and Integrating BI Governance with Corporate Governance. In *International Journal of E-Entrepreneurship and Innovation*. *4* p1–14. http://doi.org/10.4018/ijeei.2013040101 (Accessed on 2018-03-19)

[27] https://en.wikipedia.org/wiki/Corporate_governance (Accessed on 2018-03-19)

[CLARIN ERIC 2016] CLARIN ERIC. 2016. CLARIN Newsflash September 2016 | CLARIN ERIC. https://www.clarin.eu/CLARIN-Newsflash-September-2016 (Accessed on 2018-03-19)

[CLARIN ERIC 2017] CLARIN ERIC. 2017. Value Proposition. https://www.clarin.eu/value-proposition. (Accessed on 2018-03-19)

[Eckart, Hellwig, and Goosen 2015] T. Eckart, A. Hellwig, and T. Goosen. 2015. *Influence of Interface Design on User Behaviour in the VLO*. In *CLARIN Annual Conference 2015 Book of Abstracts*. https://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf (Accessed on 2018-03-19)

[Finnis, Chan, and Clements 2012] J. Finnis, S. Chan, and R. Clements. 2012. *Let's Get Real -How to Evaluate Online Success?* . https://www.keepandshare.com/doc/3148918/culture24-howtoevaluateonlinesuccess-2-pdf-september-19-2011-11-15-am-2-5-meg?da=y (Accessed on 2018-03-19)

[Goosen and Eckart 2014] T. Goosen, and T. Eckart. 2014. Virtual Language Observatory 3.0: What's New? In *CLARIN Annual Conference 2014 in Soesterberg, The Netherlands*. http://www.clarin.eu/sites/default/files/cac2014_submission_2_0.pdf (Accessed on 2018-03-19)

[Haaf, Fankhauser, Trippel, Eckart, Eckart, Hedeland, Herold, Knappen, Schiel, Stegmann, and van Uytvanck 2014] S. Haaf, P. Fankhauser, T. Trippel, K. Eckart, T. Eckart, H. Hedeland, A. Herold, J. Knappen, F. Schiel, and D. van Uytvanck. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres. In *Clarin 2014 Conference [CAC2014]*. http://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3210/file/Haaf_Fankhauser_CLARINs_virtual_language_observatory_under_scrutiny_2014.pdf (Accessed on 2018-03-19)

[Sugimoto 2017] G. Sugimoto. 2017. Number game. In *ArXiv:1706.05089 [Cs]*. http://arxiv.org/abs/1706.05089 (Accessed on 2018-03-19)

[Van Uytvanck, Zinn, Broeder, Wittenburg, and Gardelleni 2010] D. Van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardelleni. 2010. Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, M. Rosner, D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC\textquotesingle10)* (pp. 900–903)Valletta, Malta: European Language Resources Association (ELRA).