Examining Web User Flows and Behaviours in CLARIN Ecosystem

Go Sugimoto
ACDH-OEAW
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 business intelligence of web analytics to measure the impact of the aggregation services and research infrastructures in cultural heritage and digital humanities.

1 Background

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 implement objective evaluation which would determine the course of its sustainability. In this regard, Culture24 echoes the needs of web measurements within the UK museum sector (Finnis et al. 2012). Although several evaluations have been conducted for CLARIN (Eckart et al. 2015; Sugimoto 2016), their scopes are often limited to a single website and/or they focus on the technical services of the research infrastructure. For this reason, 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. 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 served as a communication and dissemination website. In addition, there are many web applications developed by the CLARIN developers such as Virtual Language Observatory (VLO)[2], Content Search Aggregator[3], and WebLicht[4]. They are useful research tools and deployed either in the subdomains of CLARIN or its partners domains. The users jump from the main website, or directly go to, those services to start a research. Regarding the VLO, it is a resource discovery 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. The users are directed to the repository of a data provider where the resources they find in the VLO search engine is stored. Although more limited, the users also navigate between the CLARIN services and the partner websites.

[1] https://www.clarin.eu/content/mission-and-strategy
[2] https://vlo.clarin.eu/
[3] https://spraakbanken.gu.se/ws/fcs/2.0/aggregator/
[4] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

As such, there are at least three major entry points to the CLARIN websites (the main website[5], the CLARIN applications, and the national centres) and the movements of the users among those websites are complex. 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 the users in the large "ecosystem" which cannot be recognised by the previous research of a single website observation.

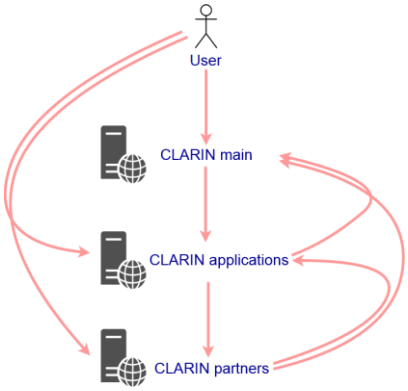

Figure 1. Multi-layered CLARIN web structure ("ecosystem") and user access

2 Analysis

The data range of this project is between February 1st 2016 and January 31st 2017, taking technical limitations and comparative studies into account (Figure 2). In order to reconcile the broad spectrum of the CLARIN's web structure, the author inspects the following websites by Piwik[6]: the main website, the VLO, the WebLicht, the Content Search Aggregation, the Discovery Service, and the Identity Provider.

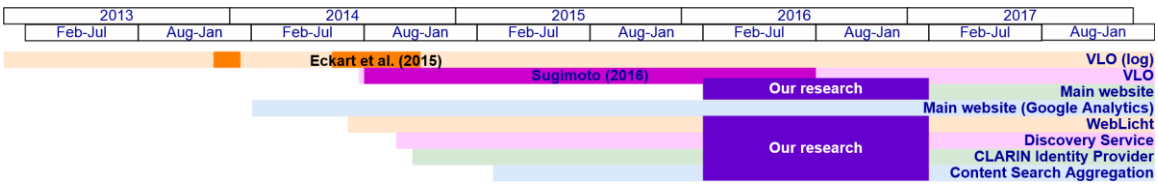

Figure 2. Research period coverage
(Available data periods are represented in light colours and research periods in dark colours)

3 VLO

3.1 Entry gate to CLARIN ecosystem (VLO)

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. transition view). 21,945 page views are recorded in the period, in which 23% are from internal webpages, 18% from search engine, 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 coincident, for example, when searching linguistic information. As for the outbound paths, 51% of the users remain in the main website, of which 12% are through to Services, 11% to Events, 8.8% to Participating Consortia, 5.5% to Clarin-

[5] 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.
[6] https://piwik.org/

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 to attract the users among other countries.

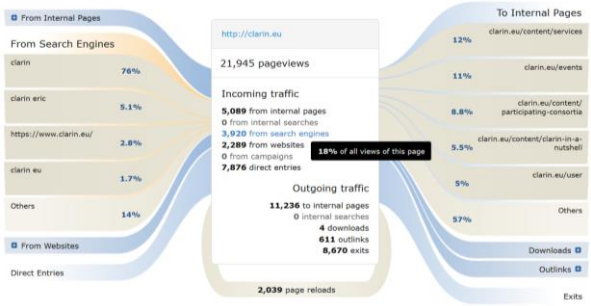

Figure 3. User flow
at the home page of the main website

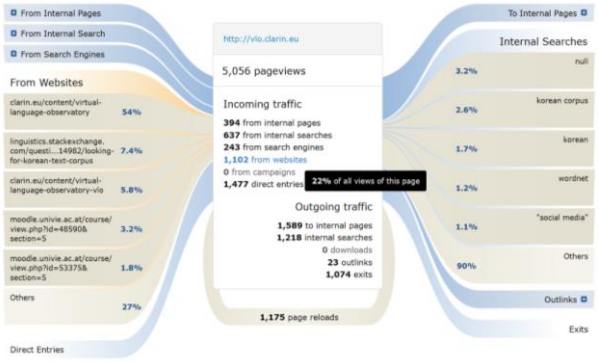

Figure 4. User flow
at the home page of VLO

From the VLO's point of view, the trend of in- and out- channels is different (Figure 4). 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, 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 as the University of Vienna offers a Moodle link to the VLO. Unlike the main website, a low number of users landed the VLO via search engine (4.8%). 29% of the users directly find the website. Regarding the outward traffic (leaving the home page), we can see the clear trend of Korean probably caused by the abovementioned stream ("korean" (1.7%) and "korean corpus" (2.6%)).

3.2 Other connection points of the ecosystem (VLO)

There is a VLO introduction page at the main website which would be one of the main gates to the VLO. 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. A part of the problem is that the individual URIs of the centres need to be checked and the use of Persistent Identifiers (Handle) 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 among others.

4 WebLicht and Content Search Aggregator

The user flow of the WebLicht is relatively simple. 70% of its visits 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 the top score of idp.clarin.eu (29%). Similarly, Germany dominates the visits by country, however, there are interests from South Korea and the United States outside Europe (Figure 5). The visit duration is substantially longer (11 minutes 57 seconds on average) than the VLO (4 minutes 18 seconds) (Sugimoto 2016). 27% spent more than 10 minutes, proving the characteristics of the data processing service. As for the Content Search Aggregator (Figure 6), the 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, the 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.

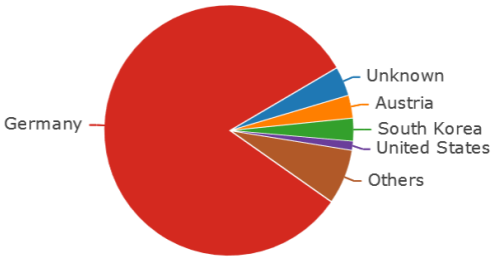

Figure 5. WebLicht visits by country

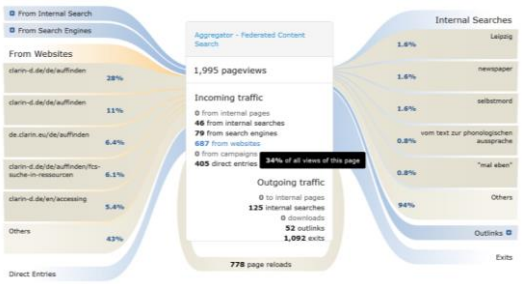

Figure 6. User flow
at Content Search Aggregator

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 the CLARIN's web space to support our previous analyses. Interestingly, both the Discovery Service (i.e. all users trying to access log-in services) and the CLARIN Identity Provider (i.e. only users with CLARIN credentials) have a large proportion of exit (65% and 82% of outbound traffic respectively). It may imply that many users give up the 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 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.

6 Conclusion

The transition view of Piwik 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 general, the existing users, characterised by direct entry and "intended access by search engine" seem to influence the statistics to some extent. It is still too early to draw a conclusion that the CLARIN websites rely on the internal community users, however, initial results are affirmative, even when they are compared to the outcomes of Sugimoto (2016) who suggested a heavy usage of the VLO by a CLARIN partner in Austria. The VLO received the most outbound traffic of the main website, mainly through its introduction page. The high volume of flow from Germany can be seen in different traffic records, but the population bias is not yet take into consideration. The impact of sudden increase of particular access such as "Korea" became easily visible from the beginning to the end of the access paths. There are also some areas where further research is needed to clarify the situation and provide correct interpretation. The preliminary results of this paper successfully displayed new insights 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. It is hoped that this small research has brought some ideas about the visitors and environments of the CLARIN's virtual ecosystem and would be a valuable contribution to the development and sustainability of CLARIN.

References

[Eckart et al. 2015] T. Eckart, A. Helwig, and T. Goosen. 2015. Influence of Interface Design on User Behaviour in the VLO. In *CLARIN Annual Conference 2015 Book of Abstracts*.

[Finnis et al. 2012] J. Finnis, S. Chan, and R. Clements. 2012. *Let's Get Real -How to Evaluate Online Success?*.

[Sugimoto 2016] G. Sugimoto. 2016. October. *Number game –Experience of a European research infrastructure (CLARIN) for the analysis of web traffic*. CLARIN Annual Conference 2016, Aix-en-Provence, France.