Number game –Experience of a European research infrastructure (CLARIN) for the analysis of web traffic

Go Sugimoto
ACDH-OEAW
Vienna, Austria
go.sugimoto@oeaw.ac.at

1 Background

The European Commission (EC) described in the principles for access to research infrastructures: "Research Infrastructures should have a policy defining how they regulate, grant and support Access to (potential) Users from academia, business, industry and public services" (European Union, 2016). Coincided with the principles, it is appreciated that CLARIN has a strong backing from linguistic community in Europe and successfully undertaken a series of projects over the last few years, demonstrating the value of academic and research infrastructure in its own right. However, despite the emphasis on the users in the principles, it is evident that there is a lack of user evaluation in CLARIN. The most recent publication (Eckart et al., 2015) reports on the user behaviour of Virtual Language Observatory (VLO), one of the flagship services of CLARIN, but it does not offer any previous literatures on the subject. Although there are some user evaluations, they seem to be limited in the form of internal community feedback (Goosen & Eckart, 2014. Haaf et al., 2014). Wynne (2015) conducted a personal analysis on a various types of users and target domains, but there seem to be several contradictions on the target users and his conclusions are a bit of overstatement without providing proper evidences to prove them. This paper is intended to present objective statistics and make assumptions and conclusions in a more scientific manner. This is a vital step for CLARIN to confront with, because the question now is whether the CLARIN community will continue developing its infrastructural services without adequately and systematically assessing its use, or not. It is also hoped that this paper could lead to more intensive user evaluation including observations, interviews, online and offline questionnaires, and focus groups. Moreover, unsatisfied with a myriad of "blind" statistical reports on web visitors and following conclusions which cannot be applied to other websites, the paper will try to shed light upon the documentation of "science of web analytics" as a subject of research, sharing and discussing the process, methodology, and experience of analyses and interpretations, which hopefully brings a few lessons for the research infrastructure and Digital Humanities field alike.

2 Open evaluation –users and transparent European research

In terms of user-driven business development, Europeana offers a good example as an e-infrastructure in the area of Digital Humanities and cultural heritage. In April 2016, it launched a beta version of Europeana Statistics Dashboard[1]. It allows the users to examine the web traffic of Europeana from 2013 to the present. It is equipped with an easy-to-use interface to view the data from different angles. Although there are a few concerns about the "privacy" or "confidentiality" of web traffic information of the content providers of Europeana, it is most likely the transparency of the cultural heritage data aggregator overshadows the disadvantages. The author believes that it can trigger positive technical incentive to improve the metadata (and associated digital objects) provided by thousands of content providers, as well as a political sentiment for the EU members to "compete" with each other. In a broader context, a new form of research evaluation is considered to be urgently needed. This is especially discussed in Open Science initiatives. For example, the EC is keen to develop a methodology to measure research quality. Expert Group on Altmetrics clearly and rightly stated (Wilson 2016): "Wider use of quantitative indicators, and the emergence of Altmetrics, can be seen as part of the transition to a more accountable and transparent research system." With this trend in mind, the analysis in this paper does

[1] http://statistics.europeana.eu/

not simply offer statistics and user evaluation, but also a hint of *Open Evaluation* of a research infrastructure in general. In other words, it is a challenge to answer such questions as how we can provide an open and fair development and services and how to measure their impact on the users? Apparently the simple web statistics cannot be compared with such complex metrics, but it would try to serve as a small forward-thinking contribution to CLARIN as a research infrastructure. Indeed, the most envisaged value of this paper is to raise awareness of the user-centric, transparent, and measurable approach for research and development in CLARIN.

3 Analysis between zero and one

Due to the limitation of space, this abstract is only able to present preliminary results, giving a flavour of the methods, process, and results of the analysis. CLARIN started to record the web traffic by Piwik[2] on June 23rd in 2014, although there are some variations of starting dates for different websites. The author decided to set a baseline analysis period from July 11th 2014 to July 10th 2016. That gives us reasonable information about the monthly and annual trend over 24 months, when recent critical development has been made.

3.1 Web applications (VLO, WebLicht, and BAS)

This section explores the traffic trend of the web applications of CLARIN. Among others, three distinct services (Virtual Language Observatory (VLO)[3], WebLicht[4], and Bavarian Archive for Speech Signals (BAS)[5]) will be examined. For example, figure 1 illustrates the two-year period of web traffic in VLO. At first glance, it is not very good news. The average unique visitors and visits per day are 16.5 and 19.6 respectively (figure 1). The highest numbers do not impress either, which are 45 and 58 respectively. There is also a decreasing trend from Year 1 to Year 2, albeit the more promotion, visibility, and development. It is obvious that more effort is needed for marketing and dissemination, as well as the reconsideration of development priorities and strategies. The top countries of access are Germany (23%), the Netherlands (11%), the United States (9%), Austria (8%), and the United Kingdom (7%). The level of engagement is also an element to measure the performance of our services and the first set of results suggests that the users spend long time in case of WebLicht (figure 2). 33 percent of visitors spent more than 4 minutes, of which 11% is accounted both 15-30 minutes and over 30 minutes. However, there can be an easy answer for the low bounce rate and long visit duration: it may be mainly due to the small number of visitors and the heavy usage of CLARIN team. More examination is needed to verify it, for example, by excluding known IPs of CLARIN partners from the analysis. It is also very useful to check what happened before and after known events such as a new release of VLO, an addition of a new CLARIN member state, and a promotion of CLARIN at a conference.

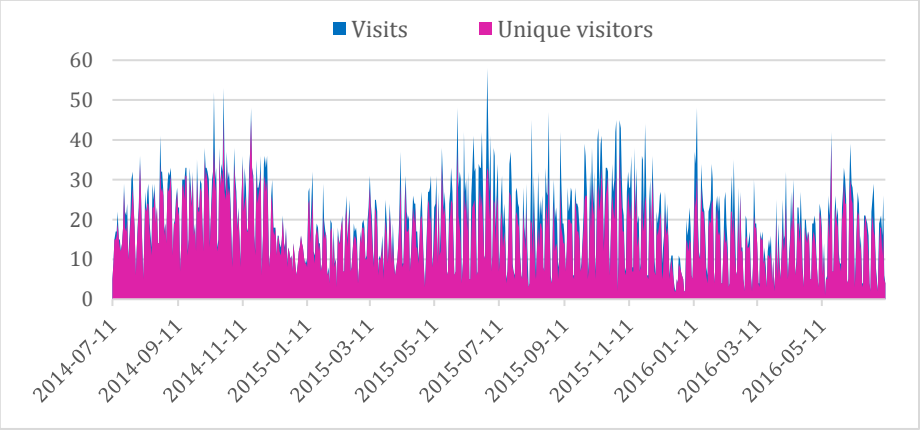

Figure 1. VLO traffic over 2 years

[2] https://piwik.org/
[3] https://vlo.clarin.eu
[4] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
[5] https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services

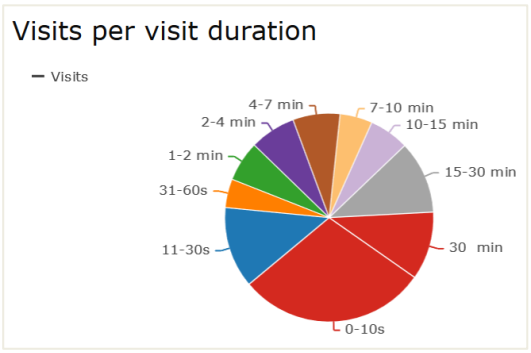

Figure 2. Visit duration for WebLicht

3.2 Identification services (Shibboleth & Co)

The Piwik also records the authentication and authorization services of CLARIN. The web statistics would be able to highlight some useful information about how the authentication is used by the users. For instance, WebLicht, Corpus of contemporary Dutch, and OpenSoNaR are the top referrers for CLARIN discovery service (figure 3), implying the degree of popularity for Dutch and German services. It would be interesting to investigate the validity of a common theory that authentication hampers access. It may be possible to examine the extent of impact of authentication on web application. Identity providers are illustrated in figure 4. Norway, Austria, Czech, and Poland are the most prominent providers, although the vast majority is CLARIN.eu. More details of those identification services will be presented as to the types of user, locations, user devices and so forth, in comparison with web applications in the section 3.1.

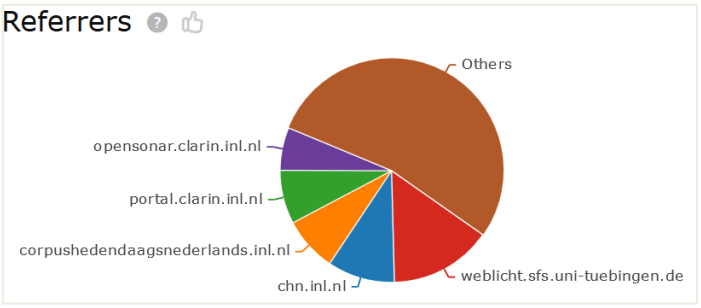

Figure 3. Top referrers for Discovery Service

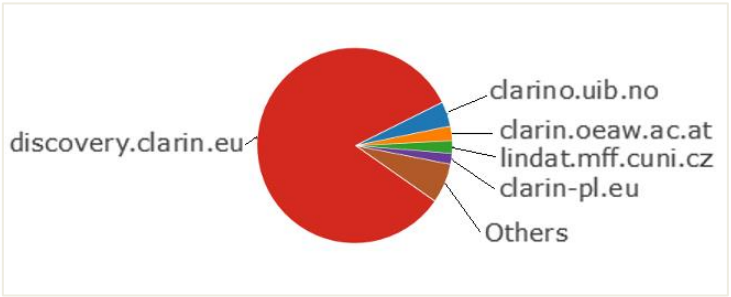

Figure 4. The share of identity providers

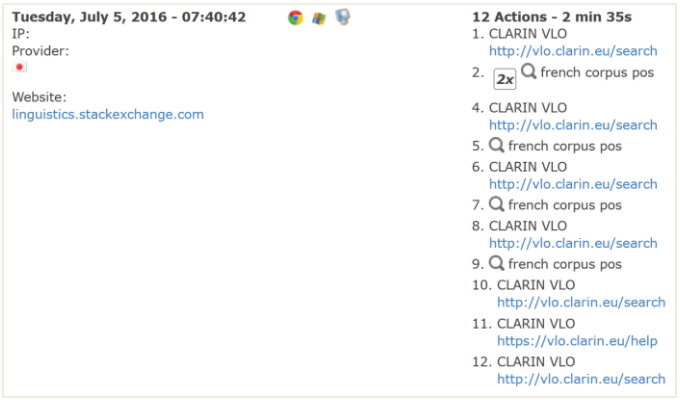

Figure 5. Visitor logs tracing user behaviours
(IPs are removed for privacy reasons)

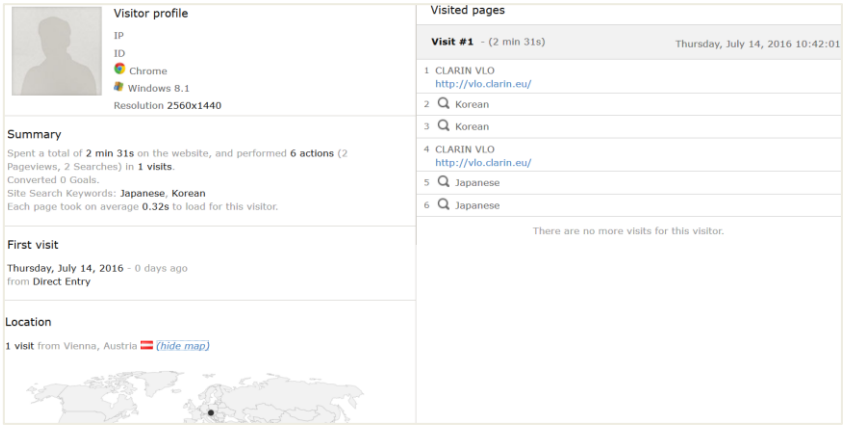

Figure 6. Visitor profile of the author with the log on the right

4 Playing a fair game? –methodological issues of web analytics

This section aims to discuss the procedural and methodological aspects of web analytics in general. It is intended to focus on "the science of our experience" on analyses, hypotheses and interpretations of website evaluation, which is relatively untouched in the studies of the internet traffic. One of the main questions of web analytics is how trustful they are and how to adequately analyse and interpret the statistics they produce. Web analytics are often a black box and most marketers blindly use them without knowing exactly what is recorded within. To this end, one experiment is made to understand how Piwik keeps tracks on the web users. Its visitor log and visitor profile function enable the author to monitor every move of a particular user, himself in real-time and the past (figure 5). An arbitrary protocol was set to execute a series of tasks and the visitor log is analysed what is recorded. For example, a protocol can be as follows: 1) type vlo.clarin.eu in the address bar of CHROME browser, 2) type search term "Hindi" in the search box, 3) click the first result of a result list, 4) click resource tab, 5) click the first resource, and 6) browse the resource on the external website and close the window. A preliminary trial indicates that the basic information such as country, operating system, and browser is correct, while the actions taken are perhaps not the same as expected (figure 6). Search terms are accurately displayed, but the tool fails to trace all the paths of the tasks. This implies a need of care for the web analysis. In a similar way, the visitor log can be anonymously investigated further to understand what a particular user and/or user group look for, and what the potential needs are. As such, if Piwik is adequately used, serious marketing research can be carried out and business and market potential can be unlocked.

5 Mind the gap between the creators and users – A little Kaizen

The current problem of CLARIN's technical development is the absence of a simple PDCA (Plan-Do-Check-Act) methodology. As far as the author recognizes, the Check is preliminary carried out by internal testing and feedback, concentrating on bug fixing and self-contained improvement. Also it is not connected to marketing strategies. This is apparently not enough. This article has presented a minimum set of user analysis, attempting to address the issues of the user-oriented development method with respect to the web traffic. In a business environment, it is critical to check the Return of Investment (RoI). Online analytics provide one of the most important Key Performance Indicators (KPIs) which are measurable over time and objective, as contrary to the more indirect success indicators such as educational impact, in order to assess the RoI. It is not the conclusion of this paper to overemphasize the needs of user analysis, rather it is hoped that it has given enough incentive for the CLARIN community to try to invest more on marketing and strategy building alongside the technical development. As the CLARIN expands and plays a leading role in the Digital Humanities, there is no doubt that user centric approach will be a major element of its operation. This paper has also tried to examine and understand the mechanism and implication of a web analytics tool, which is necessary to interpret the phenomena of the users. It hopefully enables the readers to find universal lessons instead of project-specific results of the internet traffic. In addition, this paper raises awareness of *Open Evaluation* under transparent management, which includes the sharing of user statistics. As seen in other initiatives, it is a growing concern for the credibility of CLARIN as open research infrastructure. This document is just a beginning of a little Kaizen[6] in many areas and in every level of CLARIN.

Reference

[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*. https://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf

[European Union 2016] European Union. 2016. European Charter for Access to Research Infrastructures Principles and Guidelines for Access and Related Services. http://ec.europa.eu/research/infrastructures/pdf/2016_charterforaccesstoris.pdf#view=fit&pagemode=none

[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

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

[Wilsdon 2016] J. Wilsdon. 2016. Next generation altmetrics: responsible metrics and evaluation for open science. Call for evidence. https://ec.europa.eu/research/openscience/pdf/call_for_evidence_next_generation_altmetrics.pdf

[Wynne 2015] M. Wynne. 2015. Users of CLARIN - who are they? https://www.clarin.eu/blog/users-clarin-who-are-they

[6] https://www.kaizen.com/about-us/definition-of-kaizen.html