

Examining Web User Flows and Behaviours in CLARIN Ecosystem



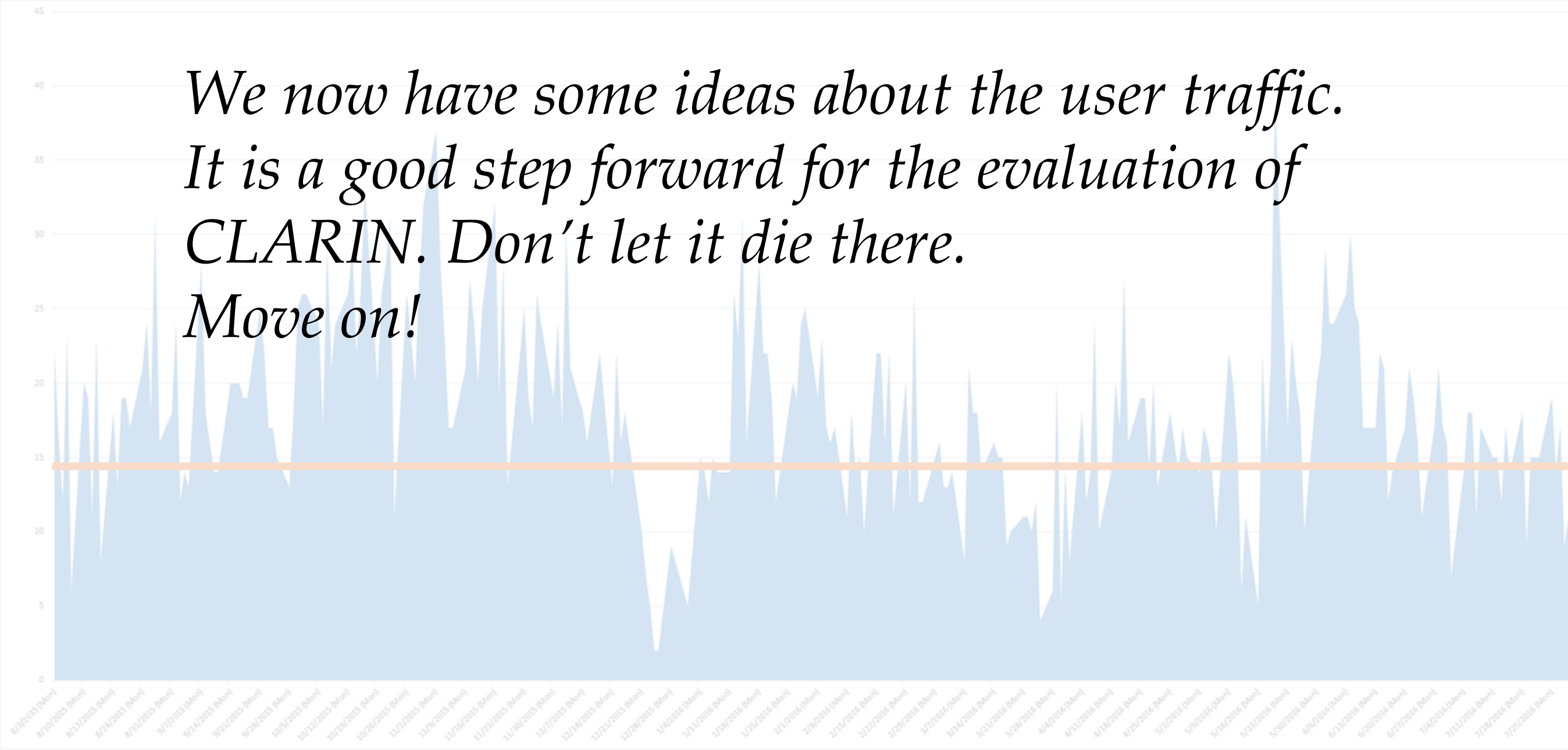
Go Sugimoto (Austrian Centre for Digital Humanities)
CLARIN Annual Conference 2017, Budapest, Hungary
2017-09-20

Agenda

1. Motivation
2. Method
3. Statistics & Interpretation
4. Conclusion

1. Number Game 2.0.x

*We now have some ideas about the user traffic.
It is a good step forward for the evaluation of
CLARIN. Don't let it die there.
Move on!*



CLARIN Virtual Language Observatory

Welcome to the VLO!

...hundreds of thousands of language resources, or continue to
...your area of interest or discover new resources.

Previous studies

- Eckart et al. (2015)
- Sugimoto (2016)

Analysis only for

- a single website
- a technical website (=VLO)

Language	⌵
Collection	⌵
Resource type	⌵
Modality	⌵

EXMARaLDA DEMO CORPUS

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

⊞ A selection of short audio and video recordings in various languages to be used for instruction or demonstration of the EXMARaLDA system.; HIAT (simplified); HIAT; free comment; suprasegmental information; accentuation/stress; English translation; Standard German translation; German translation; Englisch translation; code-switch



1

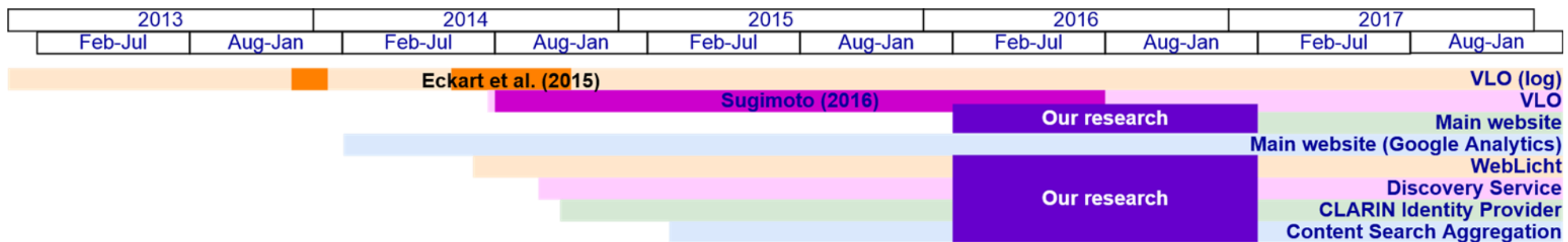


The Hamburg MapTask Corpus (HAMATAC)

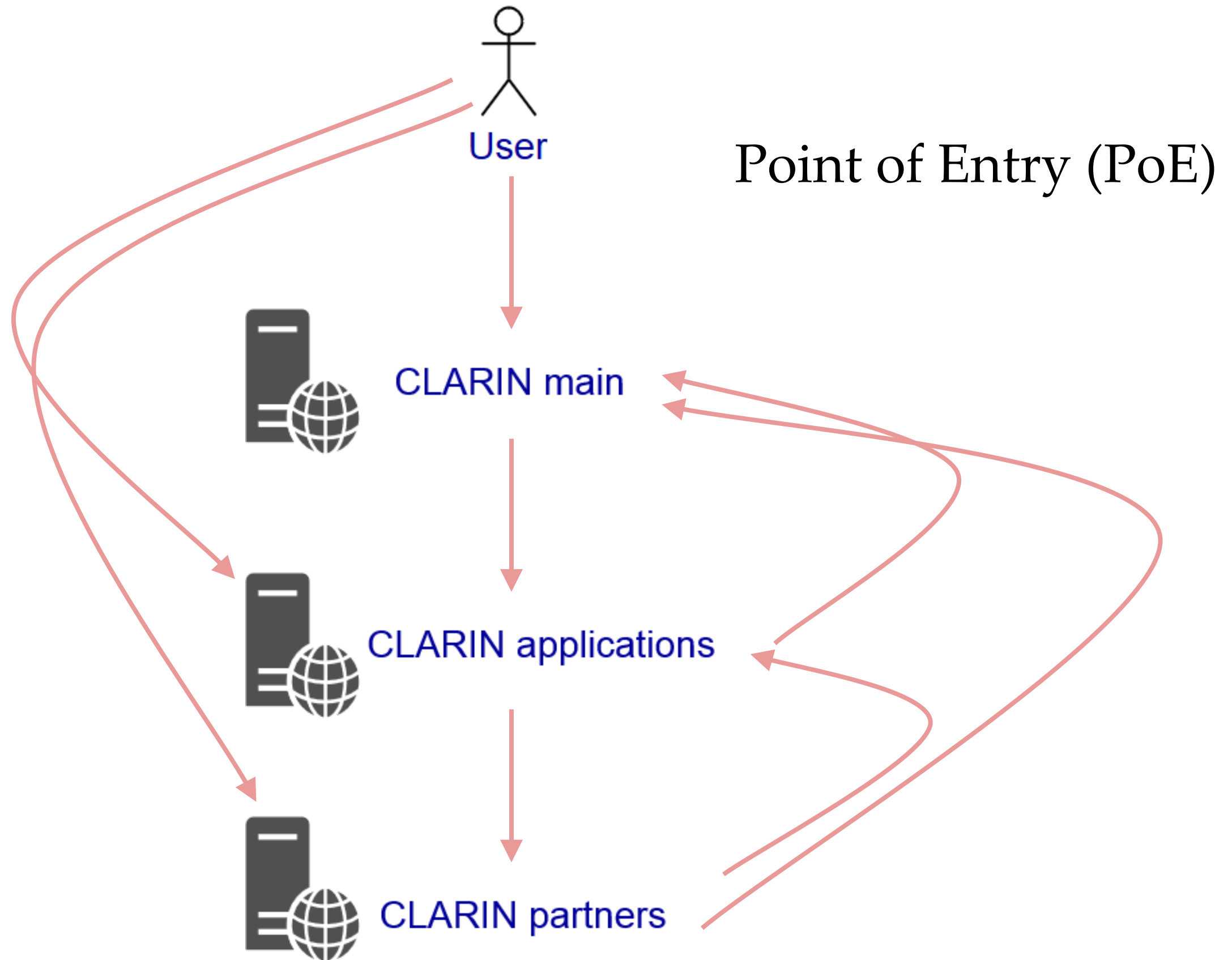


4

Web Analytics overview



2. Method -Measuring CLARIN ecosystem



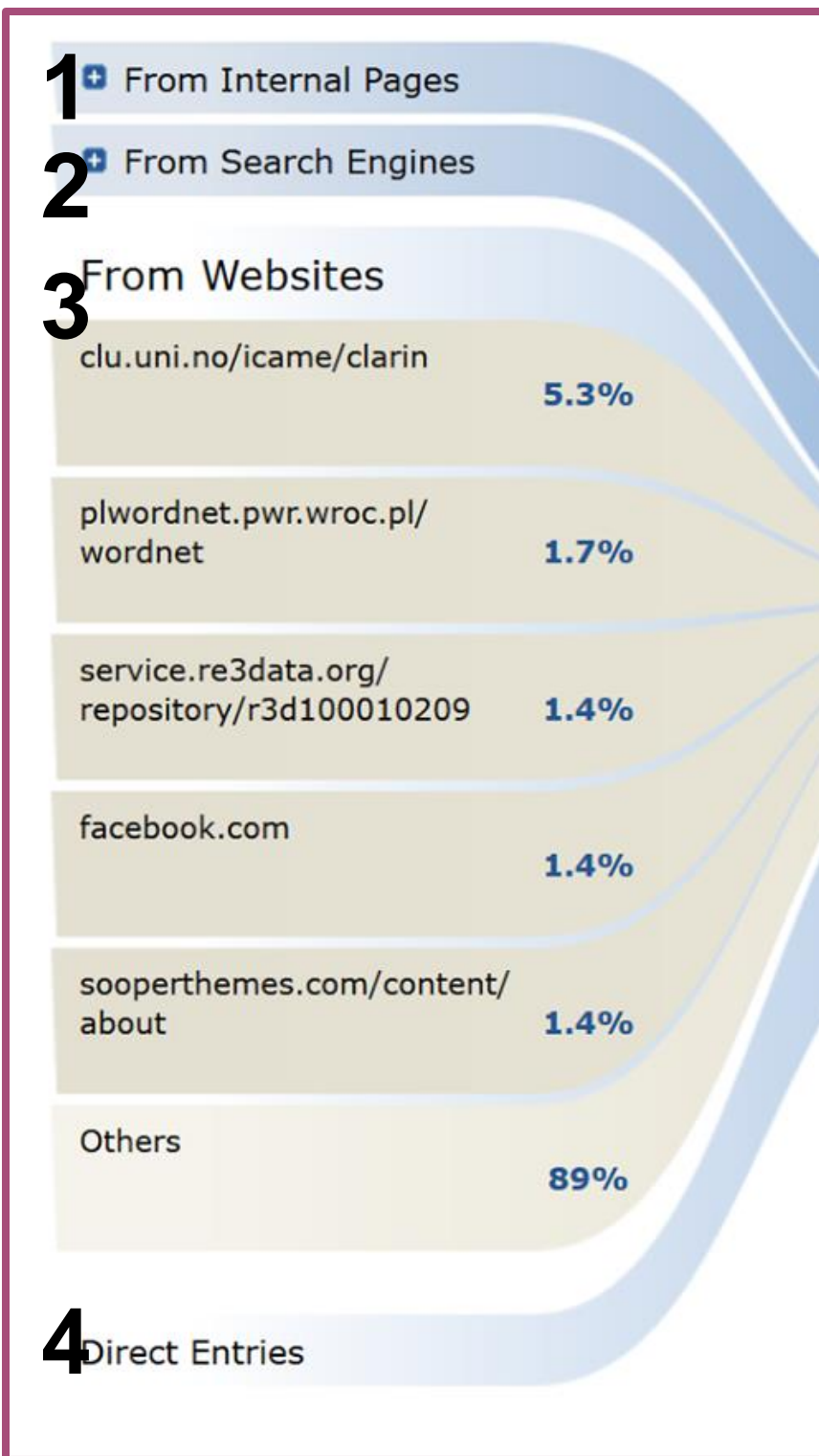
PIWIK

Open Analytics Platform

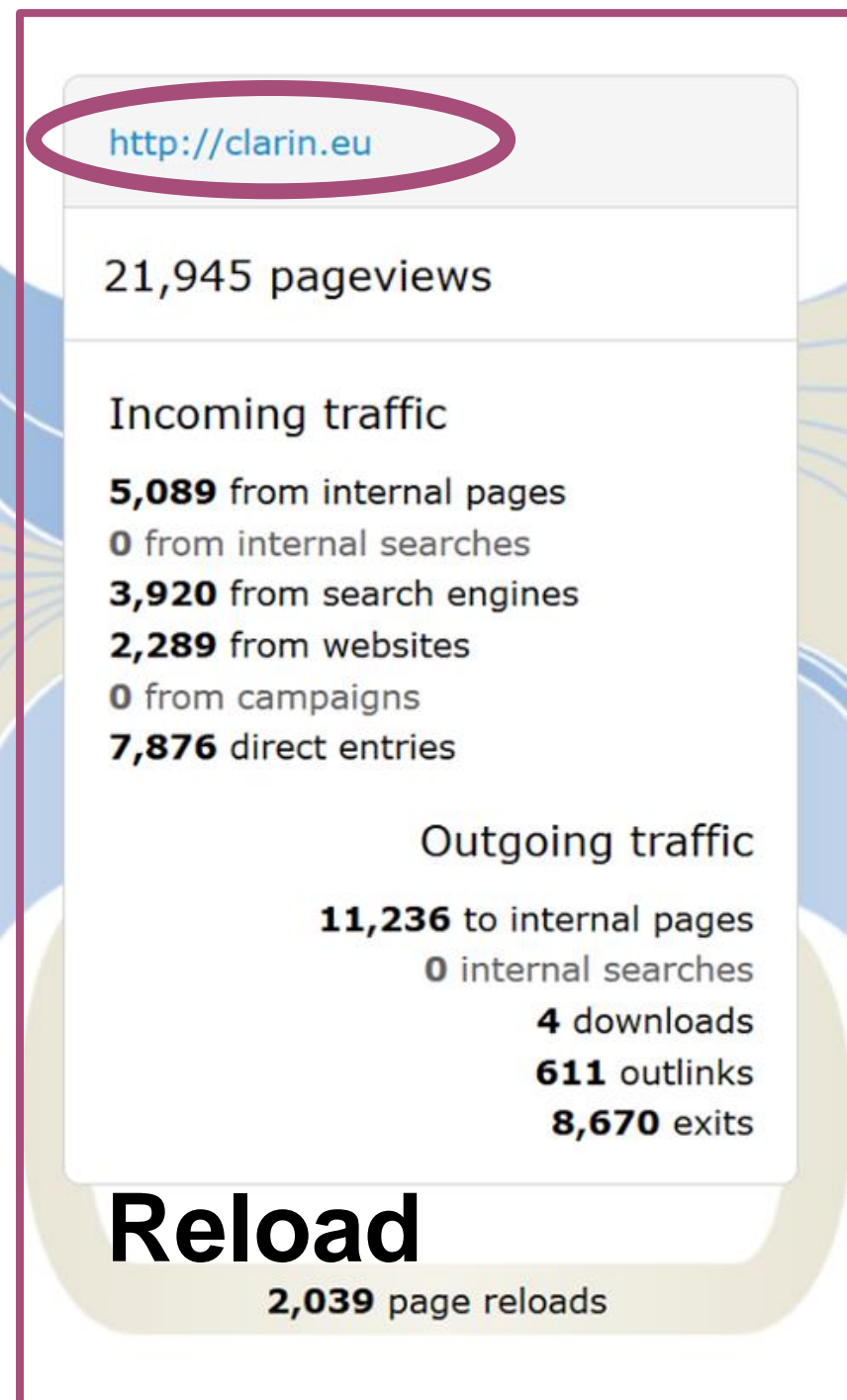
1 February 2016 - 31 January 2017

Comparative study (Traffic in and out of):

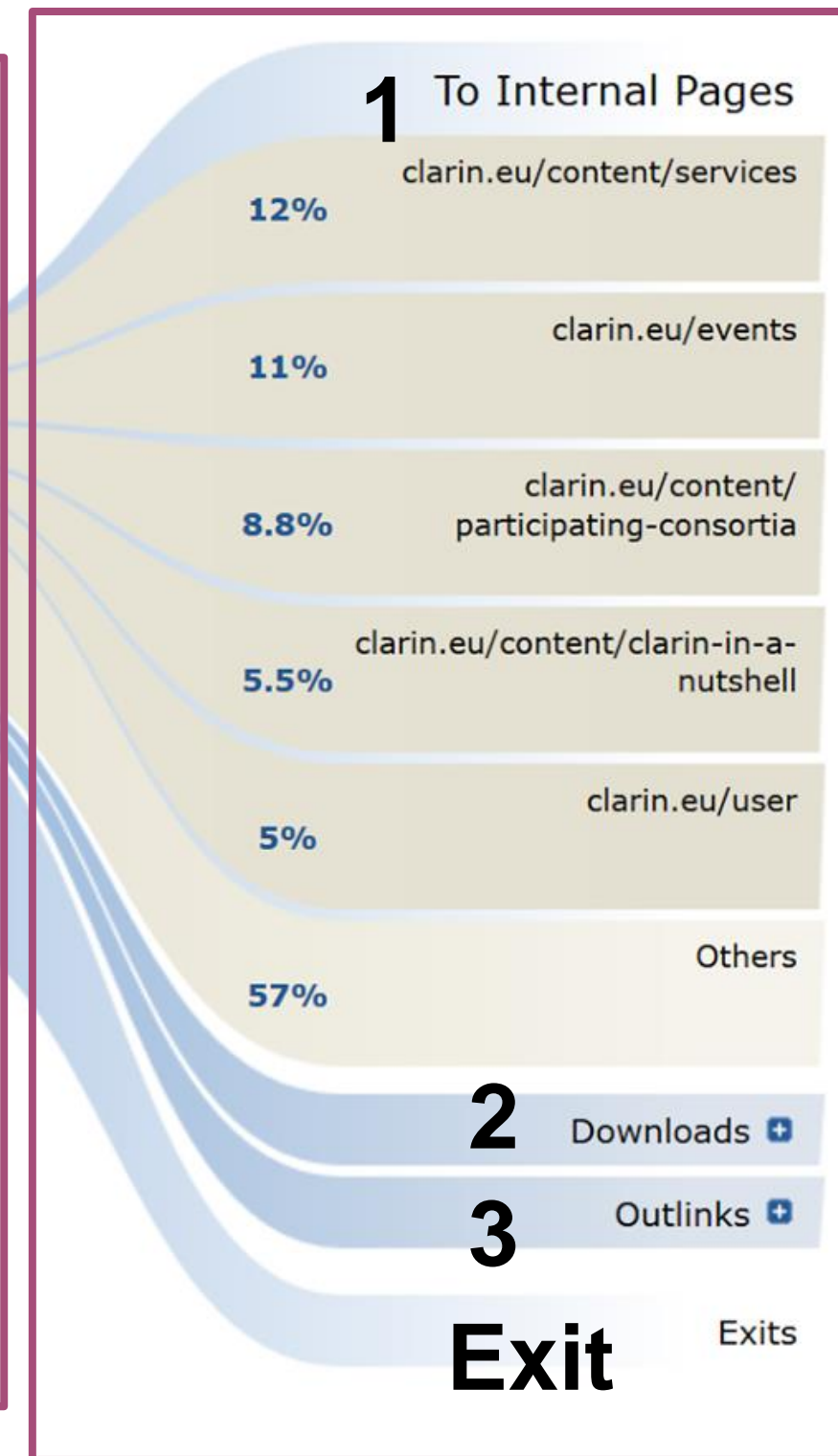
- Clarin.eu (main website)
- Virtual Language Observatory (VLO)
- Weblicht
- Content Search Aggregation
- Identity Provider
- Discovery Service



Where you were
BEFORE



Standpoint



Where you were
AFTER

ICAME corpora in CLARIN

ICAME corpora have found a new home in the [CLARIN project](#) through the Norwegian part, [CLARINO](#).

The corpora are available for academic use.

The corpora are available in the [Corpuscle program](#) (login from the top line).

Users from many universities can log in with their ordinary user id through the [eduGAIN](#) or [CLARIN Service Provider Federation \(SPF\)](#). Go to the [Corpuscle](#) home page and choose eduGAIN or CLARIN SPF from the top login line and search for your university.

Norwegian users can use Feide login through eduGAIN (if you don't find your institution on the login page contact Knut.Hofland@uni.no).

If you are not able to use eduGAIN or CLARIN SPF register for an [ClarInIdP](#) account and be manually approved.

[OpenIdP](#) is another choice, but this may be terminated on short notice.

At the moment the following corpora are available for searching (most also for downloading through the "Overview" option in the menu to the left for each corpus):

- The Brown family (Brown, LOB, FLOB, Frown, BLOB and BE06) prepared at Lancaster (BLOB and BE06 not for downloading)
- FLOB and Frown with original POS tagging
- ACE (Australian Corpus of English)
- COLT (Corpus of London Teenage Language)
- Helsinki Corpus of English Texts
- Helsinki Corpus of Older Scotts
- Helsinki CEECS (Corpus of Early English Correspondence Sampler)
- London-Lund Corpus

5.3% of referrers

<http://clu.uni.no/icame/clarin/> ->

<http://clarin.eu>

Outgoing Traffic

51% Internal pages

12% Services

11% Event

8.8% Consortia

5.5% Nutshell

5% Users (probably log-in)

2.8% Outlinks

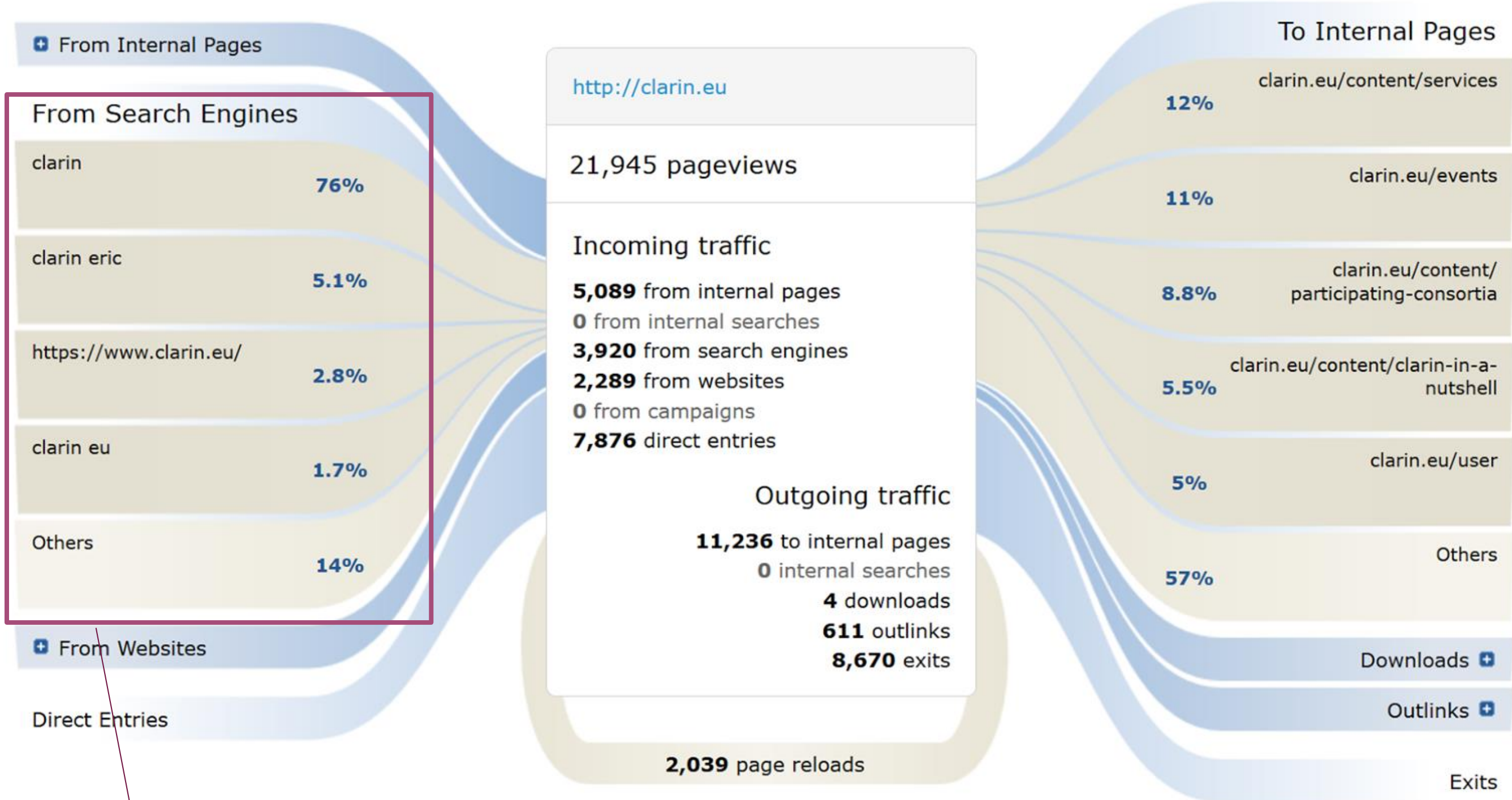
30% VLO

3.8% CLARIN-D

Some noises

40% Exit

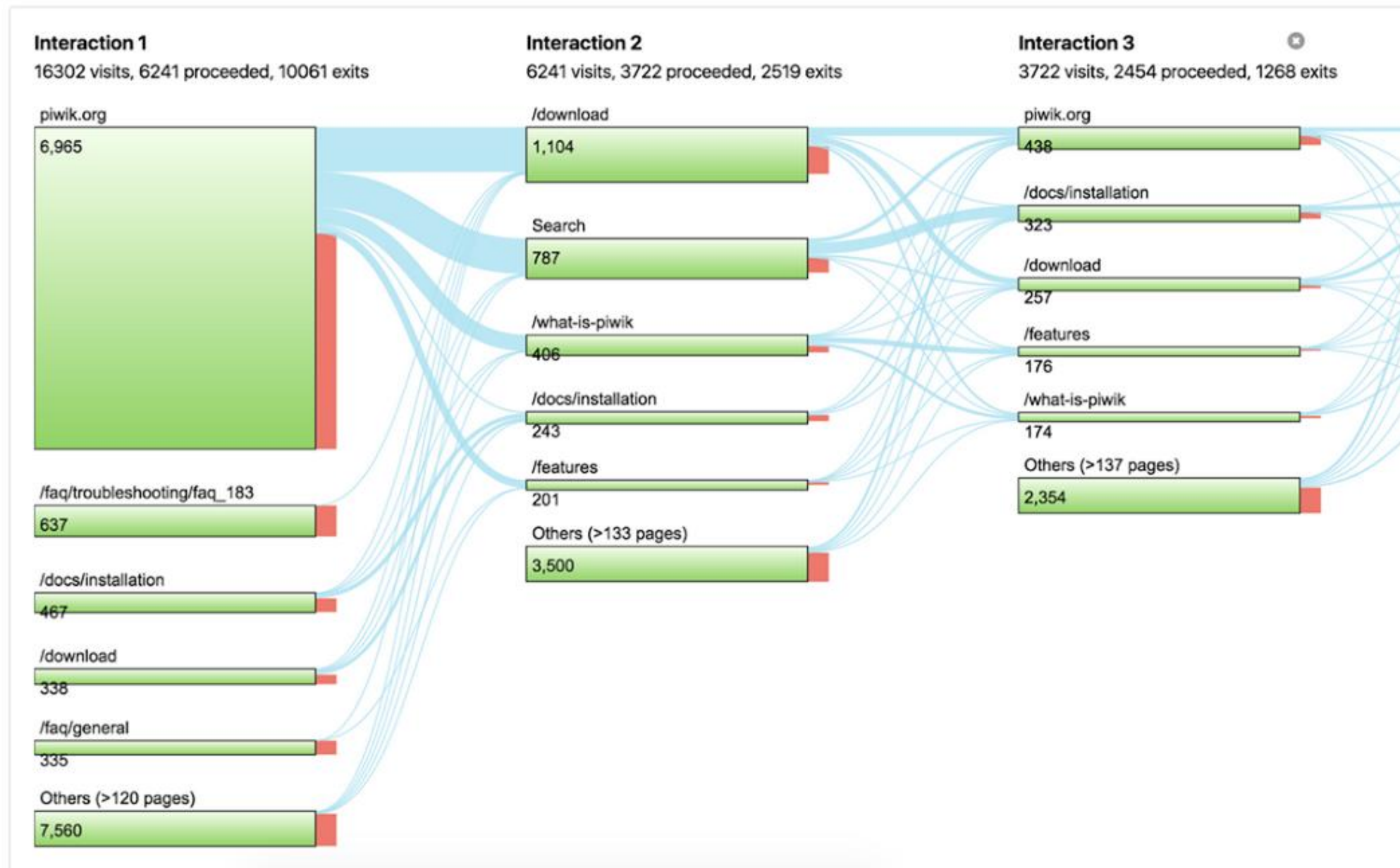
Depends on what is put (or promoted) on the home page which changes frequently



Do they know CLARIN already?

100% - (Reload and Internal pages) = 14817 (external access)

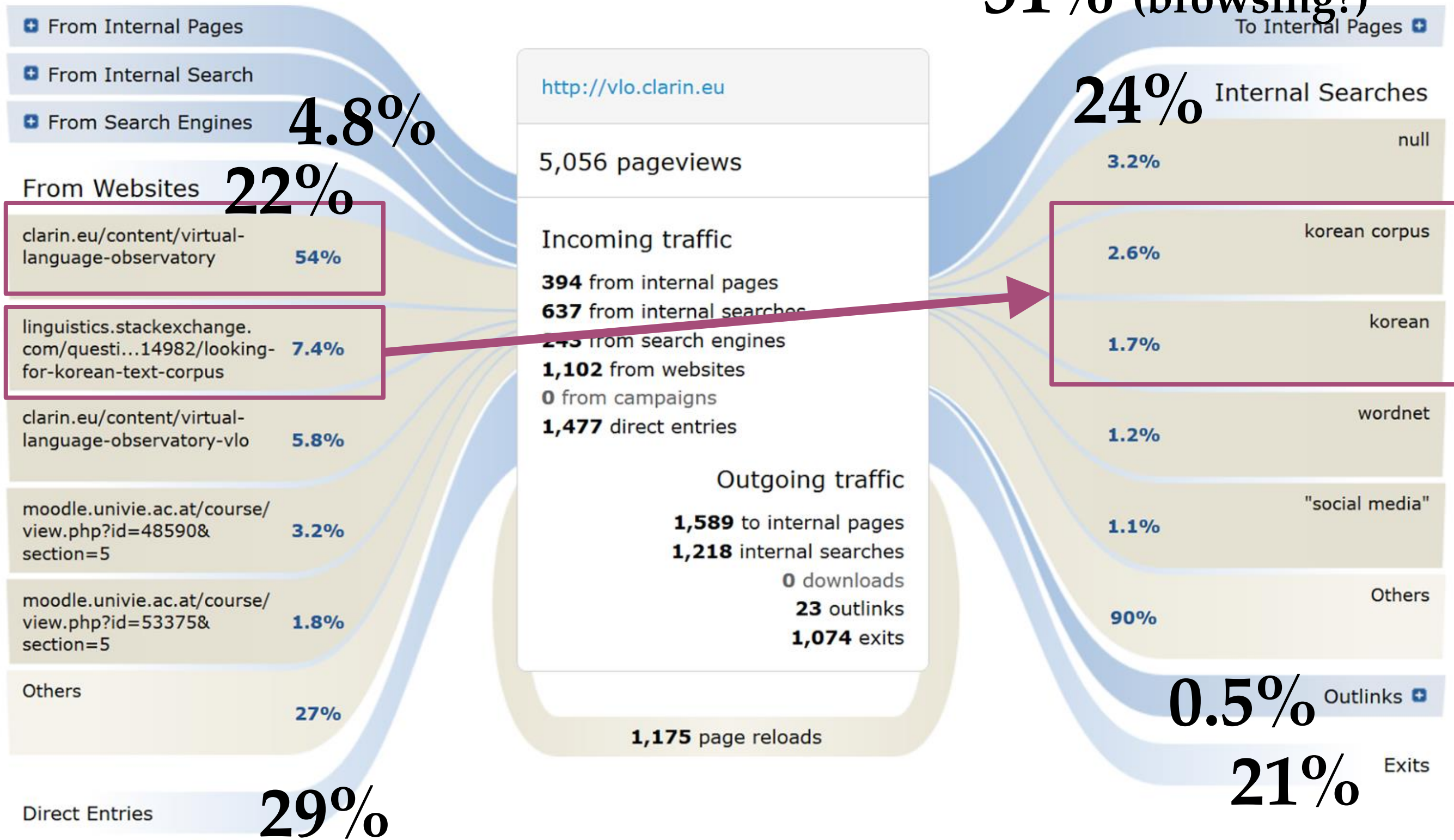
Search engine with CLARIN + Direct Entry = 8028 (existing users?)
 = 54.2% minimum + other channels



User Flow Plug-in (€79 /year) is more powerful than Transition Reports
 It can check the position a page was viewed how often and how users navigated through your entire website. Multiple-user paths can be tracked.

	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%

Top 10 outlinks from the CLARIN.EU domain

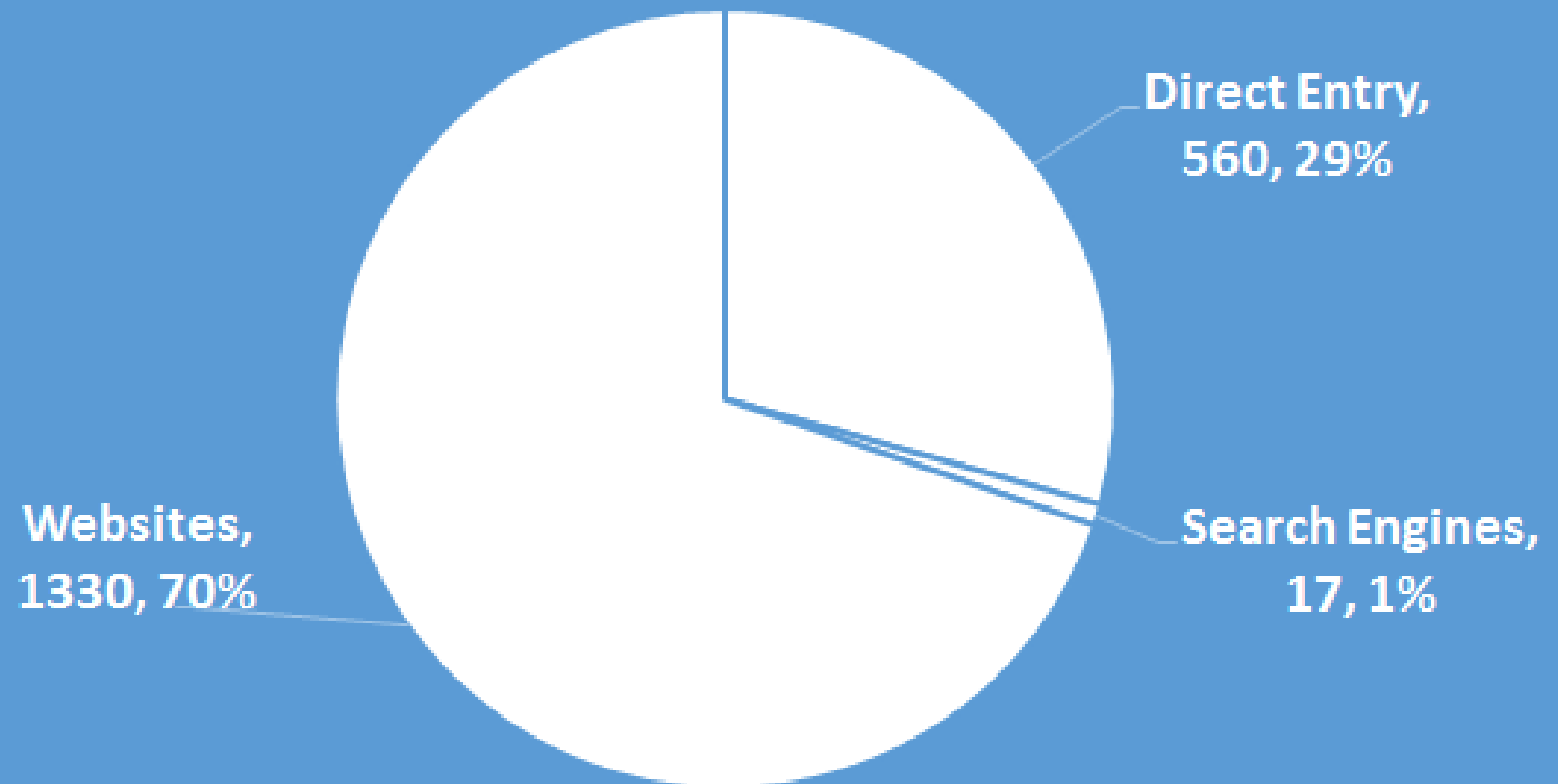


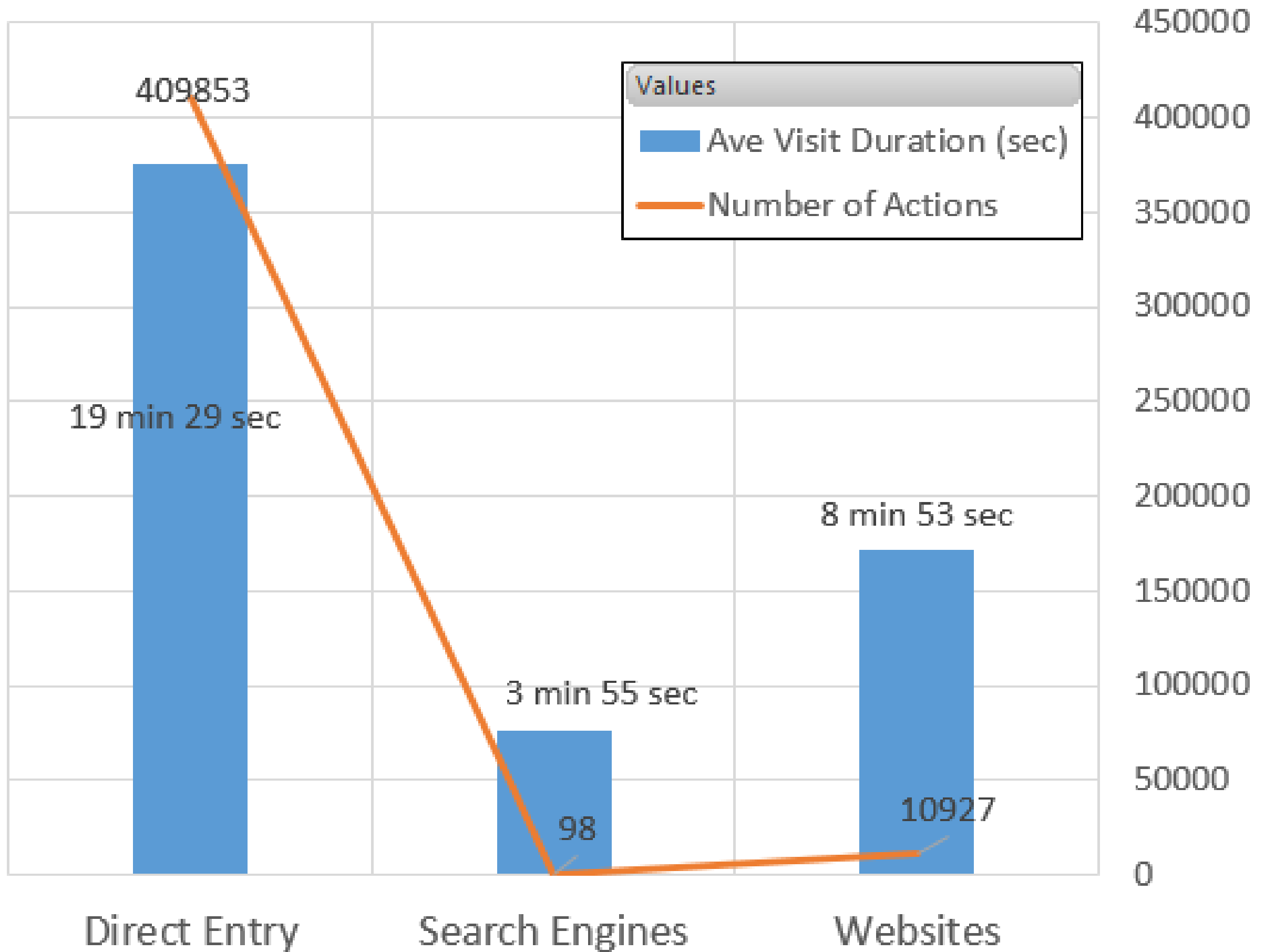
At VLO home page

Ranking	URL	Unique Clicks	Percentage
1	hdl.handle.net	297	41.4%
2	infra.clarin.eu	118	16.4%
3	www.clarin.eu	96	13.4%
4	clarinportal.informatik.uni-leipzig.de	20	2.8%
5	www.sil.org	15	2.1%
6	clarin-pl.eu	12	1.7%
7	urn.fi	11	1.5%
8	diglib.hab.de	10	1.4%
9	clarinws.informatik.uni-leipzig.de:8080	8	1.1%
10	metashare.ut.ee	8	1.1%

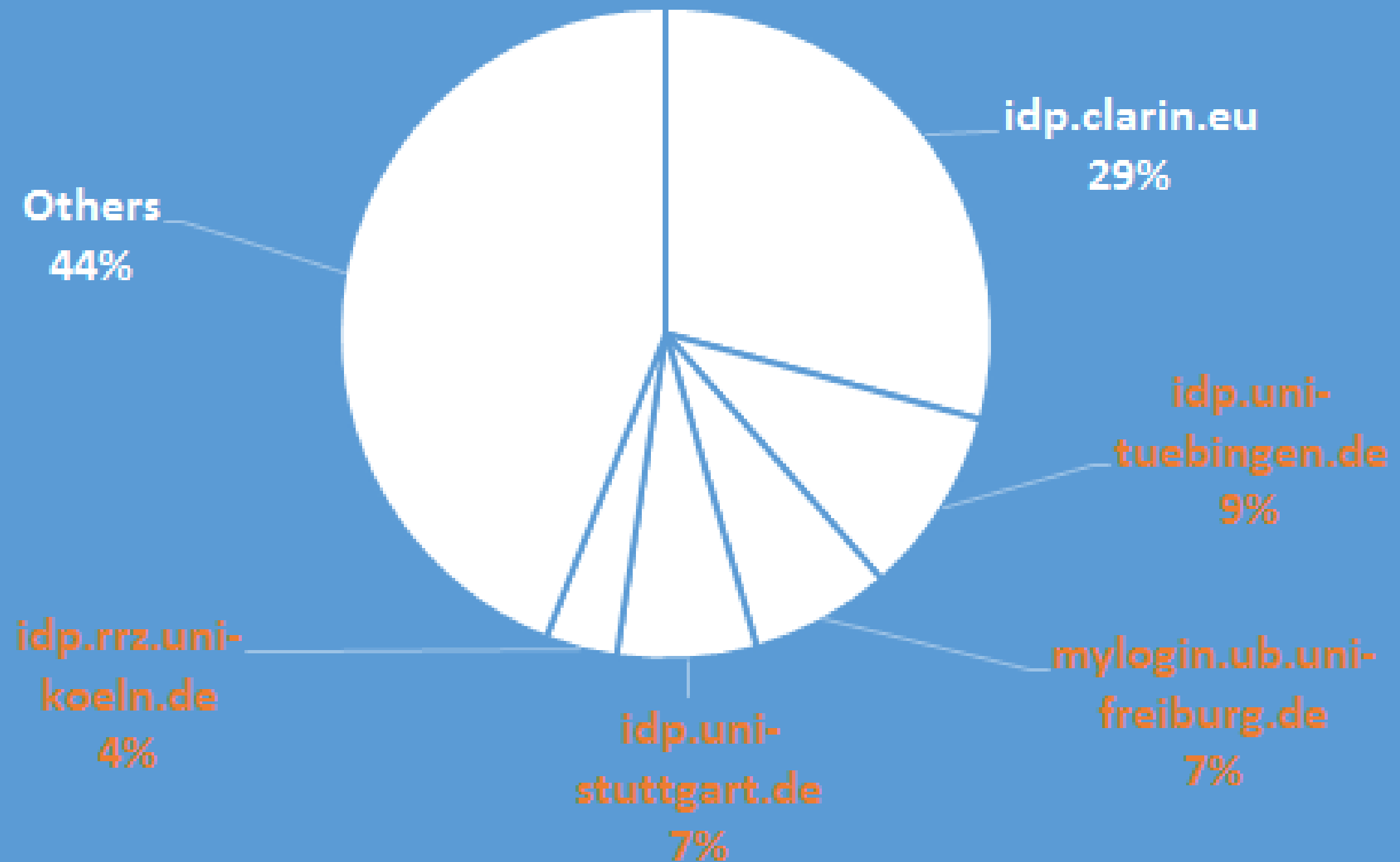
Top 10 outlinks from the VLO domain

WEBLICHT - TYPES OF INCOMING TRAFFIC














WEBLICHT - WEB REFERERS



Limitations and Dirty Jobs for Analyses

- Duplicates, flatening, and development with CMS
<https://www.clarin.eu/content/factsheet-clarin-plus> (287 visits)
<https://www.clarin.eu/node/4213> (235 visitis)
- Handle PIDs have no clue for the content, while URLs include clues (domain names etc)

PAGE URL	PAGEVIEWS	▼ UNIQUE PAGEVIEWS
 content	42,002	28,722
 /index	22,434	16,019
 event	10,079	7,524
 news	4,961	3,779
 /events	3,209	2,242
 search	2,998	2,233
 node	2,905	2,089
 /4213	235	158
 /3760		

Open Transitions

See what visitors did before and after viewing this page

4. Conclusion

- Piwik (transition view) -simple yet powerful tool for user assessment
- CLARIN - complex web ecosystem with many PoE
- Good starting point for the better understanding of the users
- Tricky to analyse & interpret data. Don't use default data blindly (as most people do...)
- Open Evaluation! **Agile** Evaluation!

Future work

- Combine with other surveys (questionnaire etc)
- Use/buy plug-ins for better analyses
- Web Analytics to synch with Outreach & Dev team
- Case study with the institutions in the consortium (1. Piwik data exchange to track the user flows at the partner site. 2. Distinguish the CLARIN users by IP address from non-CLARIN users etc)
- Social media analytics

Your ideas & collaboration welcome!

Köszönöm!

Original title of this paper: “Donau in Blue” was abandoned with regret. The concept was:

- User flow = Danube river
- The Blue Danube (Strauss II) + Rhapsody in Blue (Gershwin)
- Vienna to Budapest

Go.Sugimoto@oeaw.ac.at

Extra information

1.4% of referrers. No link! (website changed? noise?)

<http://www.sooperthemes.com/content/about> -> <http://clarin.eu>

Try our products for free: trysooperthemes.com – No credit card needed.

SooperThemes

Drupal Themes

Drupal Modules

Download

Pricing

Help Center

Contact

Log In

Join Now To Download

About SooperThemes

1

Drupal Theme

10

Years Experience

1123

Happy Customers

100%

Dedicated

“clarin”

“digital language resource”

CLARIN ERIC |

<https://www.clarin.eu/> ▼

CLARIN makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities

Services |

<https://www.clarin.eu/>

CLARIN port
resources in

Portable S

<https://www.clarin.eu/>

The Clarin b
comfortable.

Clarín | D

<https://www.clarin.eu/>

Define clarín

Searches related to clarín

diario clarín de hoy últimas noticias

diario nacion

diario perfil

diario infobae

clarín empleos

diario cronica

clarín clasificados

pagina 12



CLARIN ERIC |

<https://www.clarin.eu/> ▼

CLARIN makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities ...

Language

www.springer.com

Language Res
Academic Seal

Brave New

<https://books.google.com/books?id=...>

Robert J. Blake
Technology and
Hawai'i, Nation

Searches r

digital language lab software

digital language definition

stanford digital language lab

language lab software for schools

computer language used in internet

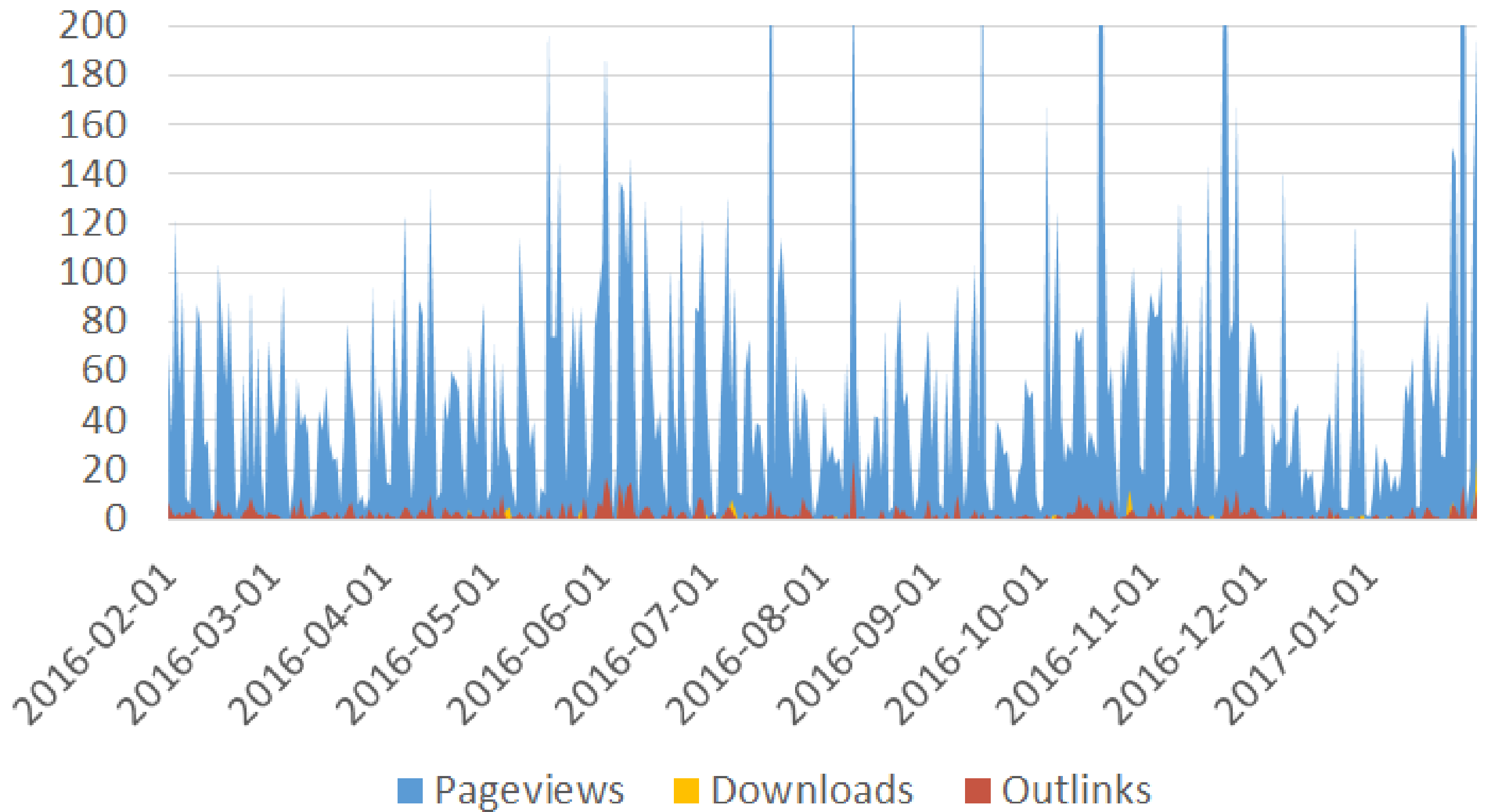
lathrop library

binary language

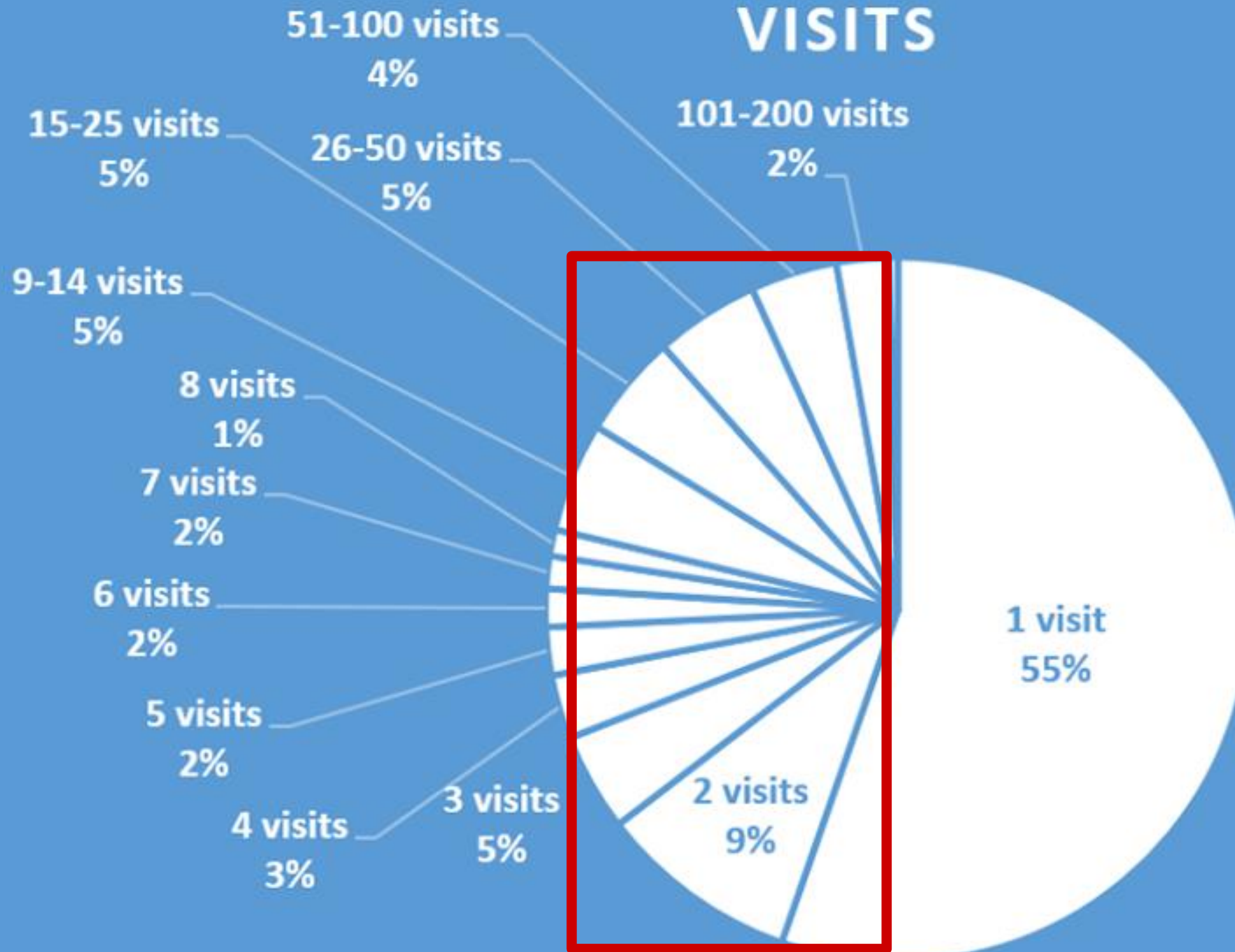


- Search Engine Optimisation (SEO)!
- Existing users not new users? (likely)
- Are they CLARIN members? (likely)

VLO pageviews, outlinks, downloads



VISITS



Dashboard

Visitors

Actions

Pages

Entry pages

Exit pages

Page titles

Site Search

Outlinks

HTML <title>

GOOD

Factsheet: CLARIN-PLUS | CLARIN ERIC

522

327

43%

00:00:53

28%

0.57s

BUT

CLARIN ERIC |

11,560

8,240

42%

00:00:43

49%

0.36s

CLARIN ERIC | CLARIN ERIC

10,891

7,766

46%

00:00:48

56%

0.37s

More questions than answers?

Is CLARIN addressing

- the community of CLARIN, linguistics, (Digital) Humanities, or Cultural Heritage, etc?
- Europe or also outside Europe?
- Researchers, (higher) Education, Citizen (science) etc?

Did the users find and use information (in general, resource, tool)?

Why do the users come and go?