Anna H. Turner

PUBLIC INTEREST IN DATA SURVEILLANCE AND DATA LEAKS BEFORE AND AFTER SNOWDEN

GOOGLE BIG DATA IN CROSS NATIONAL PERSPECTIVE

TECHNICAL REPORT

pracownia wydawnicza

Anna H. Turner

PUBLIC INTEREST IN DATA SURVEILLANCE AND DATA LEAKS BEFORE AND AFTER SNOWDEN

GOOGLE BIG DATA IN CROSS NATIONAL PERSPECTIVE

TECHNICAL REPORT

pracownia wydawnicza

Warszawa 2017

Anna H. Turner

The Graduate School for Social Research (GSSR)
at the Institute of Philosophy and Sociology
of the Polish Academy of Sciences

Public Interest in Data Surveillance and Data Leaks
Before and After Snowden
Google Big Data in Cross National Perspective
Technical Report

Warszawa 2017

This project was supported by the Cross-National Studies: Interdisciplinary Research and Training Programme (CONSIRT) established by the Polish Academy of Sciences and The Ohio State University

Projekt graficzny

Alina Wiszenko-Zabrowarny

ISBN 978-83-942170-5-1

Wydawca

Pracownia Wydawnicza Andrzej Zabrowarny

TABLE OF CONTENTS

INTRODUCTION 5
1. TOOL DESCRIPTION 8
2. COUNTRIES 13
3. DATA DESCRIPTION – KEYWORDS 15
Keywords characteristics 15
Keywords analysis 16
4. SEARCH INDEX 19
5. FINAL DATA BASE AND ANALYSIS DESCRIPTION 21
6. FINDINGS 23
Global Overview 23
Public Interest in Data Leaks and Data Surveillance 24
Public Interest in Data Leaks and Data Surveillance – month by month 25
Public Interest in Data Leaks – month by month 26
Public Interest in Data Surveillance – month by month 27
Global Overview Split by Continents 28
Public Interest in Data Leaks and Data Surveillance 29
Public Interest in Data Leaks 30
Public Interest in Data Leaks – month by month 31
Public Interest in Data Surveillance 32
Public Interest in Data Surveillance – month by month 33
Continents Overview 34
Africa 37
Public Interest in Data Leaks and Data Surveillance 37
Public Interest in Data Leaks and Data Surveillance – month by month 38
Public Interest in Data Leaks – month by month 39
Public Interest in Data Surveillance – month by month 40
African countries 41
Public Interest in Data Leaks and Data Surveillance 41
Public Interest in Data Leaks 43
Public Interest in Data Surveillance 45
Asia 47
Public Interest in Data Leaks and Data Surveillance 47
Public Interest in Data Leaks and Data Surveillance – month by month 48
Public Interest in Data Leaks – month by month 49
Public Interest in Data Surveillance – month by month 50

1. TOOL DESCRIPTION

To conduct analysis of public interest I have used Google Keyword Planner.

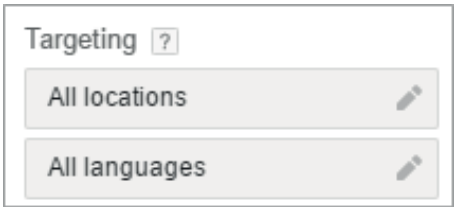
What is Google Keyword Planner

Keyword Planner is a free online tool[6] that enables users to check what and how often people are searching in Google's search engine. Usually this tool is being used by media agencies for advertising purposes, such as Paid Search or Google Display Network campaigns.[7] At the time of conducting our research Google Keyword Planner was freely accessible for all users who had Gmail accounts. Now it is also required to have an active Google AdWords[8] account (Google AdWords is a platform to set up "pay-per-click" online advertising in Google Network) and spend money on campaigns on a monthly basis, understandably this can be a serious limitation for many researchers interested in using the tool.

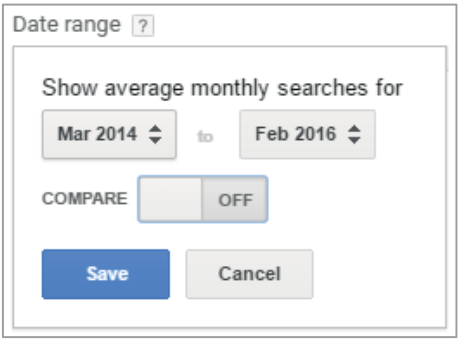
How Google Keyword Planner works

First, we start with what Google calls "Targeting". We choose location and language. We can narrow the location to country, city, province, state, region and county, but we can also conduct analysis at a global level by choosing the "All locations" tab.

[6] https://adwords.google.com/KeywordPlanner
[7] https://support.google.com/adwords/answer/2999770?hl=en)
[8] https://www.google.co.uk/adwords/

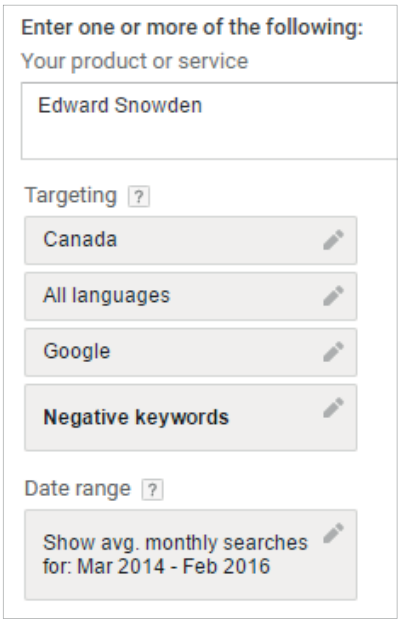


Second, we set up dates. We can check only the last 24 months and within this time we can also check monthly data or compare month to month. Assuming we would be searching from Feb 2016, then the last month we can check would be March 2014.



Third, we choose words we want to check – we call them search terms, queries or keywords – and these might not be only single words like "WikiLeaks", but also phrases or sentences, for example: "data protection" or "how to protect my privacy online".

Once we have decided when and where we want to check what people are searching for, we need to start with some ideas for keywords. Google Keyword Planner shows us data for specific keywords we started with, for example: we want to know how often in Canada people searched for "Edward Snowden" between March 2014 and February 2016; we type all the information in the Keyword Planner interface:



In result, we get information that our keyword "Edward Snowden" was searched on a monthly avg. 22 200 times.

The "average monthly searches" stands for the average number of times people have searched for the exact keyword based on targeting selected.[9] I explain below, in greater detail, about average vs. exact number of searches.

The additional value of this tool is that we can also get data on similar keywords, that people in Canada were searching for, within this period of time:

[9] https://support.google.com/adwords/answer/3022575?ctx=tltp

Keyword (by relevance)	Avg. monthly searches ? Mar 2014 - Feb 2016
snowden	8,100
where is edward snowden	260
edward snowden interview	260
edward snowdon	260
snowden news	210
who is edward snowden	210

We can get a maximum 800 similar keywords for each keyword we type. This functionality is very useful in building a database of keywords; otherwise it would be difficult to predict/guess what people typed in Google.

Methodological challenges with using Google Keyword Planner

Google Keyword Planner was designed to be a marketing tool rather than to be used in scientific analysis and as a result has significant limitations. The main issues are:

1. We do not know how data is being collected, and what algorithm is used for the Keyword Planner tool.
2. We do not have any demographic data: who is searching (no gender, age, occupation, marital status etc.) and we do not know why they are searching – what is their motivation.
3. Considering Google search engine processes on average over 54 500 search queries every second[10] (which translates to over 4.7 billion searches per day), it is probable that we omit a certain number of keywords in the analysis, even if Google returns 800 similar keywords for our one.

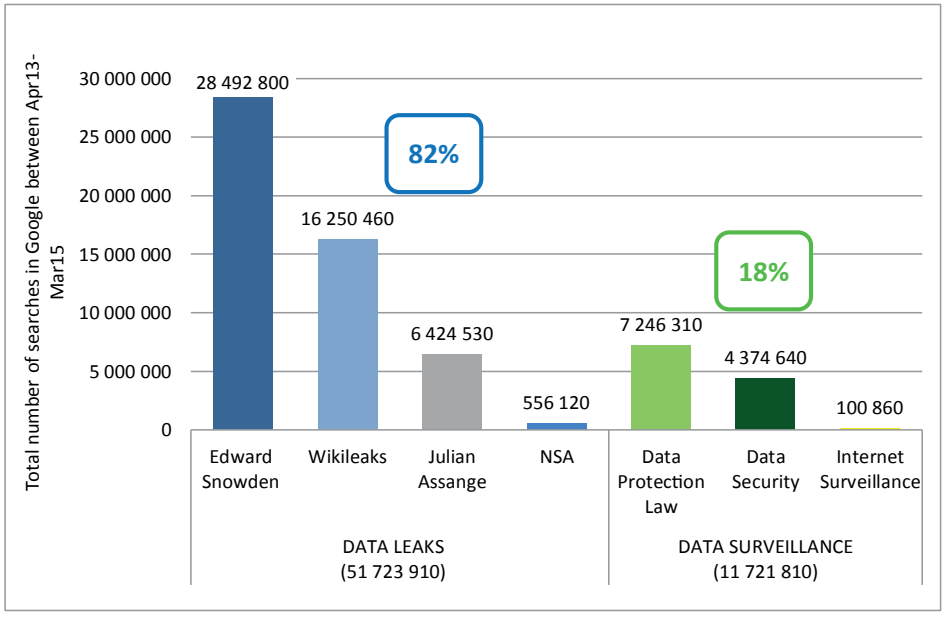
[10] http://www.internetlivestats.com/google-search-statistics/

4. There are certain non-Latin languages for which Keyword Planner shows no data, and in the research these were: Albanian, Armenian, Bengali, Belorussian, Bosnian, Bulgarian, Croatian, Georgian, Icelandic, Macedonian, Nepali, Thai, Vietnamese – as a result, analysis could not be conducted in any of these languages.
5. We also do not know the exact number of searches, as Google shows only the average number of searches, but also uses "buckets" to group keywords by traffic volume – the higher the number of searches the bigger the gap between the buckets. After analysing the data I found that Google has 82 "traffic buckets" which are logarithmically proportioned, these are: 10, 20, 30, 40, 50, 70, 90, 110, 140, 170, 210, 260, 320, 390, 480, 590, 720, 880, 1000, 1300, 1600, 1900, 2400, 2 900, 3600, 4400, 6600, 8100, 9900, 12 100, 14 800, 18 100, 22 200, 27 100, 40 050, 49 500, 60 500, 74 000, 90 500, 110 000, 135 000, 165 000, 201 000, 246 000, 301 000, 368 000, 450 000, 550 000, 673 000, 823 000, 1 000 000, 1 220 000, 1 500 000, 1 830 000, 2 240 000, 2 740 000, 3 350 000, 4 090 000, 5 000 000, 6 120 000, 7 480 000, 9 140 000, 11 100 000, 13 600 000, 16 600 000, 20 400 000, 24 900 000, 30 400 000, 37 200 000, 45 500 000, 55 600 000, 68 000 000, 83 100 000, 124 000 000, 151 000 000, 185 000 000, 226 000 000, 414 000 000, 506 000 000, 923 000 000, 1 120 000 000, 3 760 000 000. If the keyword has 74 000 average monthly searches, this does not mean this is the exact number of searches; 74 000 is between 60 500 and 90 500 (which means that the average number of searches for the keyword is closer to 74 000, but is still between 60 500 and 74 000, and 74 000 and 90 500), which gives us a 30 000 potential search difference in a single month. The higher the number of searches the greater potential inaccuracies.

Public Interest in Data Leaks and Data Surveillance

The chart below presents public interest in Data Leaks and Data Surveillance between April 2013 – March 2015 in total, in all 116 countries included in the survey, measured using total number of searches in Google. We can compare public interest between categories and topics they include.

Figure 6.1. Public Interest in Data Leaks and Data Surveillance globally



Between April 2013 and March 2015 – there was a total of 63 445 720 searches; with 82% on Data Leaks (51 723 910 searches) and 18% on Data Surveillance (11 721 810 searches). Topic wise, the highest public interest was in *Edward Snowden*, 45% of all searches, much greater than any other topic.

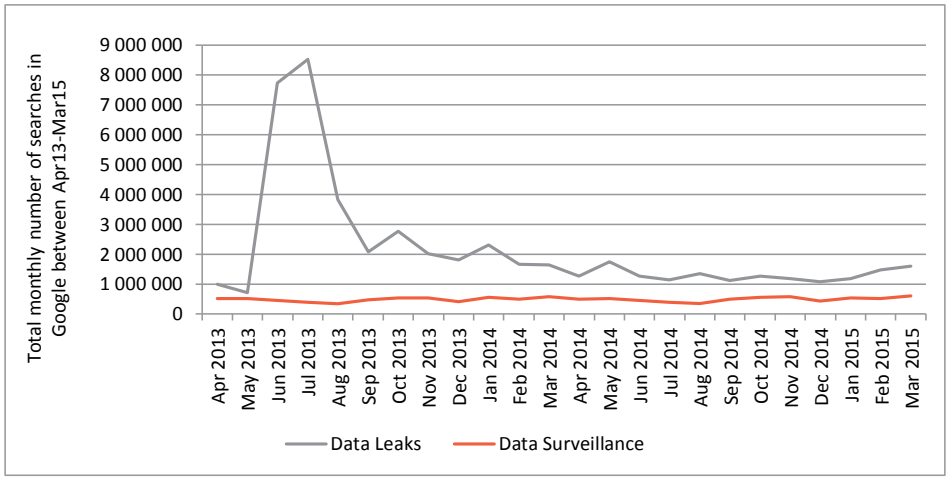
Within Data Leaks category, the highest public interest was on the subject of *Edward Snowden*, with 28 492 800 searches (55%), followed by *WikiLeaks,* with 16 250 460 searches (31%), *Julian Assange,* with 6 424 530 searches (12%) and *NSA,* with 556 120 searches (1%).

Within Data Surveillance category, the highest public interest was on the subject of *Data protection law,* with 7 246 310 searches (62%), followed by *Data Security,* with 4 374 640 searches (37%), and lastly *Internet Surveillance,* with 100 860 searches (1%).

Public Interest in Data Leaks and Data Surveillance – month by month

The chart below presents public interest in Data Leaks and Data Surveillance, in monthly intervals, in all 116 countries included in the survey, measured using total number of searches in Google. We can observe how public interest changes over the course of 24 months and compare it between the categories.

Figure 6.2. Public Interest in Data Leaks and Data Surveillance globally – month by month



Between April 2013 and March 2015, there was public interest in both categories: Data Leaks and Data Surveillance, however it differed.

Specific to the Data Leaks category was a rise in public interest in June and July 2013 – at the time of Snowden's revelations appearing in The Guardian and The Washington Post – from 711 030 searches in May to 7 732 940 in June and 8 525 420 in July.

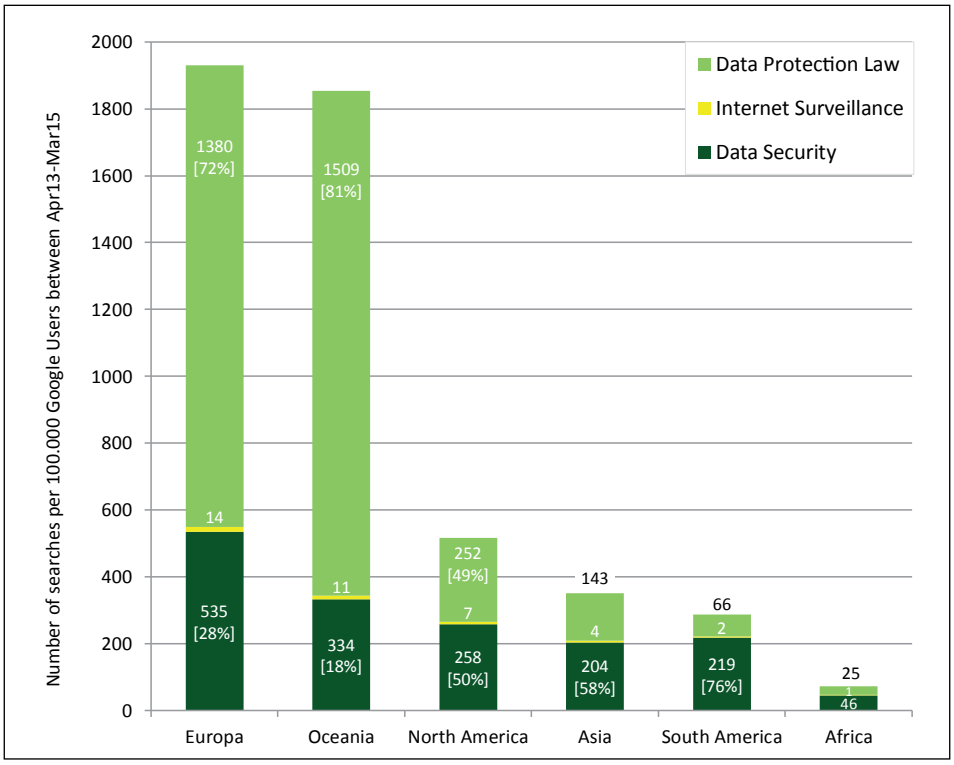
As of August 2013 public interest began to decrease, with small peaks over the following months, but in general remained stable over time.

Public interest in the Data Surveillance category was much lower and more stable during the analysed period.

Public Interest in Data Surveillance

The chart below presents public interest in Data Surveillance only, between April 2013 – March 2015 in all 116 countries included in the survey, measured using Search Index (number of searches in Google per 100.000 Google users). We can compare public interest between 3 topics within Data Surveillance category at each of the continents and between them.

Figure 6.8. Data Surveillance Continental Search Index



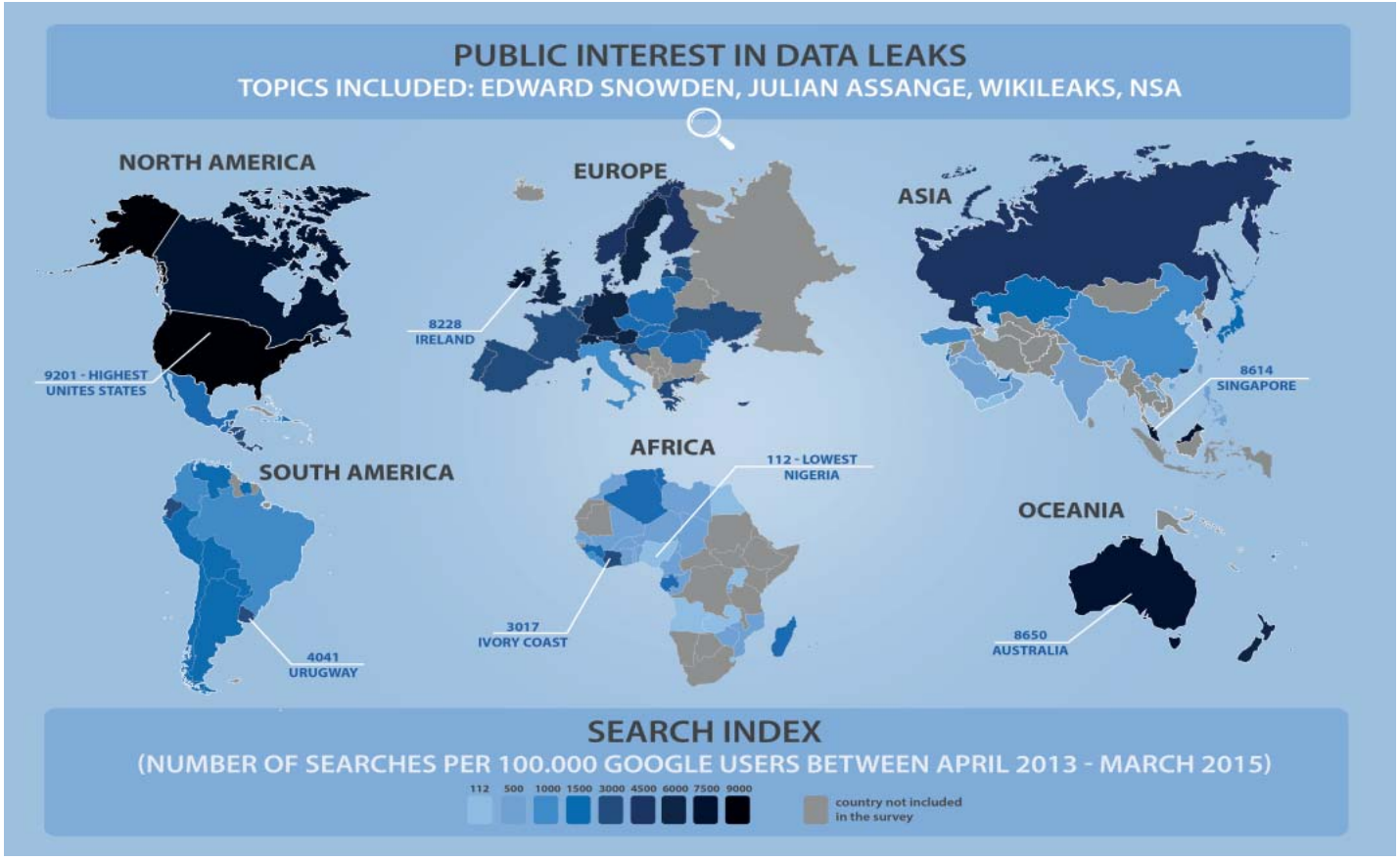
In a detailed view of Data Surveillance we see that in Europe and Oceania the highest public interest was in *Data protection law* topics, followed by *Data security* topics.

This is in contrast to Asia, South America and Africa, where public interest in *Data security* topics was higher than in *Data protection law* topics.

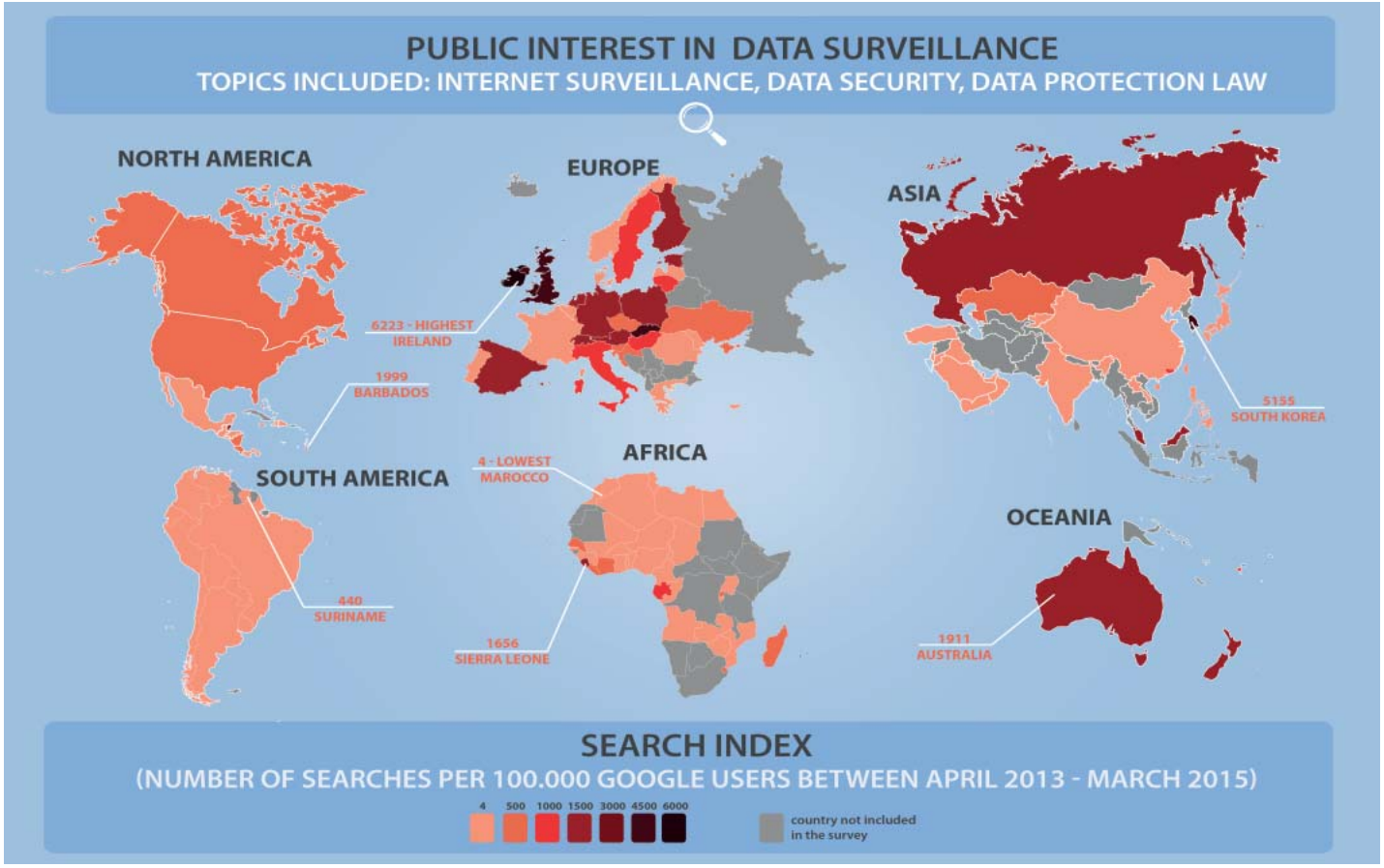
In North America, public interest in *Data protection law* and *Data security* was almost equal.

Across all continents there was relatively low public interest in *Internet surveillance*.

Infographic 6.1. All countries – Data Leaks Country Search Index



Infographic 6.2. All countries – Data Surveillance Country Search Index



Anna H. Turner is a Ph.D. candidate in the Graduate School for Social Research at the Institute of Philosophy and Sociology of the Polish Academy of Sciences. She is currently completing her dissertation on understanding the role of economic, political and social macro determinants in differentiating public interest in the topics of surveillance. Her research interests include sociology of internet, information society, privacy and data protection in the context of online surveillance and methodology of research with use of Google data.

ISBN 978-83-942170-5-1