

Discovering Data-Driven Stories: A Case Study

Läsardrivna narrativ i datavisualiseringar: En fallstudie

Carl-Johan Backman
KTH Royal Institute of Technology
Gapminder Foundation

Narrative visualization is a young and emerging field, driven mainly by data journalists. For this reason, most data stories available today are author-driven. However, with the rise of interactive visualizations the possibilities for creating reader-driven stories have become apparent.

In this thesis, we present a straightforward prototype, AsylKoll, built to support the articulation of reader-driven stories about Swedish immigration during 2015. We test its ability to support reader-driven stories by performing two user-studies based on the Think Aloud Method. In particular, we evaluate the prototype along the dimensions of reader engagement and learning. We find that user-centric data and various effects, such as transitions and mouse-overs, have a positive impact on reader engagement. In addition, we find that typical tasks such as extracting extremes and making comparisons are very important for users to gain insight and learn from the data. Foremost, this thesis shows the potential that simple, interactive visualizations have to make people engage and gain insights from data.

Additional Key Words and Phrases: Narrative visualization, data visualization, think aloud method.

Att skapa faktabaserade narrativ med hjälp av datavisualiseringar är någonting som blir allt mer vanligt idag. Utvecklingen drivs framför allt av datajournalister och av den anledningen är det typiskt sett författar-drivna historier som berättas. På senare tid har det dock blivit allt lättare att utveckla avancerade, interaktiva visualiseringar och det har öppnat för skapa faktabaserade narrativ drivna av läsaren istället.

Läsar-drivna narrativ är det som vi utforskar i den här studien. Med hjälp av en prototyp som vi byggt, AsylKoll, som visar statistik från asylinvandringen till Sverige under 2015, undersöker vi vad som krävs av en visualisering för att användare ska kunna härleda sina egna faktabaserade historier från den. Vi kollar i synnerhet på hur man kan uppmana användaren att interagera med verktyget samt vad som krävs för att användaren ska lära sig från datan.

Verktyget testas genom två användarstudier med 'Think Aloud'-metoden. I studien finner vi att data centrerad kring användaren och olika typer av effekter, såsom transitioner och mouse-overs, påverkar användarens vilja att interagera med visualiseringarna positivt. Vidare finner vi även att typiska funktioner som exempelvis möjligheten att snabbt hitta extremvärden samt att kunna göra olika jämförelser av data, är viktigt för lärandet.

Additional Key Words and Phrases: Narrative visualization, data visualization, think aloud method.

1. INTRODUCTION

Prior to any proper decision, big or small, we need information. In particular, we need fact-based information, to gain an understanding of the world to which the consequences of the decision applies. However, while we have information in abundance today, it is often hard to tell the fact-based information from the deceptive

noise. There is a subtle difference between data-driven and data-manipulated stories that has to do with intent and transparency. Given access to raw data is transparent and the only explicit intent is empowerment of individuals.

In Sweden, one topic where this difficulty is apparent, is the public debate regarding the refugee situation. It has been a widely and increasingly debated topic in recent years [Delin 2016]. During the ongoing refugee crisis in Europe, Sweden is one of the countries which has taken the greatest responsibility in offering people fleeing from war and terror a safe place to stay. In fact, Sweden received more refugees per capita in 2015 than any other European country. During the same year, more than 160 000 refugees applied for asylum in Sweden, comparable to almost 2% of the total Swedish population [Agency 2015]. However, the situation has also generated a heated debate and a polarized population. Those who are in favor of welcoming more refugees versus those who want to close the borders. As a consequence of this heated debate, misleading information stemming from doubtful sources frequently arise on social media and alternative media.

In this thesis, we aim to create and test a transparency tool that puts in the hands of individuals, interactive and visual access to the raw data regarding refugee migration towards Sweden in 2015. We report the design process and the results from pilot studies with seven Swedish citizens.

1.1 Background

The decisions made by elected politicians in Sweden are inevitably shaped by the public opinion. Moreover, the decisions made regarding this specific issue, immigration, affect the fates of families fleeing from war. Hence, the importance and severity of the debate to be carried out with correct facts and figures cannot be underestimated. Therefore, the first step when participating in such a debate should be to obtain a fact-based view of the current situation. Luckily the data about the refugee situation in Sweden is not difficult to obtain. It is publicly available on the Swedish Migration Agency's web page [Agency 2015]. However, the data is published in table format through excel sheets and PDFs and, because of that, hard for the general public to analyze and process.

For this reason, we have created AsylKoll¹ (Vizylum in English), a tool for exploring the datasets published by The Swedish Migration Agency. It is a web-based tool built on HTML, CSS and Javascript. The focus has been to create a tool easy to access and use for non-experts. The tool will be described in greater detail in Section 3.

1.2 Motivation

The work presented in this thesis is motivated by the increasing interest and use of information visualization to convey stories about

¹<https://people.kth.se/~cjba/thesis/>

data. This growing interest and practice of narrative visualization is most notable among journalists but lately also in the scientific world [Eccles et al. 2008; Ma et al. 2012]. While research in information visualization, historically, has focused on data analysis and exploration there has been a slight shift in recent years [Kosara and Mackinlay 2013]. Presentation and communication have attracted more attention and research about narrative visualizations is becoming more established. Despite that, there is a lack of studies that evaluate if the actual effects stories have on a reader are the same as those the author intended. Most of the current research focuses instead on how the visualizations are designed and framed. Another insufficiently researched subject is that of *reader-driven stories*, as defined by [Segel and Heer 2010]. Most narrative visualizations constructed today are made by journalists who have already decided on a story to tell. In that sense, narrative visualizations are currently *author-driven*, to a large extent.

However, in this age of information the amount of data in our societies is continuously increasing and while the data can be relevant to the public, only a fraction will be analyzed and presented as stories by journalists. For that reason we believe that reader-driven stories can play an important role in spreading a fact-based world view. By providing tools for the general public, which can then be used to explore data relevant to them and let them find stories of their own.

Yet another important motivation behind this study is its use for *Gapminder Foundation*, where this thesis was conducted [Gapminder 2016]. Gapminder is currently developing a framework for building visualization tools, called *Vizabi* [Foundation 2015]. Gapminder strives for spreading a fact-based world view and tries to achieve this by making data on global development easily accessible to the general public. It wants people to explore the data by themselves, using tools such as *Vizabi*, and by that gain a better understanding of the world. In other words, Gapminder wants readers to derive their own stories from the data that the foundation gathers. Hence, the insights drawn from this thesis will be valuable in the continuing development of *Vizabi*.

Taking all the above into account, *AsylKoll* was constructed as a narrative visualization intended to support the articulation of reader-driven stories, by non-experts. We have evaluated the tool through user-studies in focus groups. In particular, we wanted to explore reader-driven stories along the dimensions of reader engagement and learning.

Since reader engagement and learning, or memorability, are desired outcomes of many narrative visualizations this research should be valuable to a large fraction of the community involved with creating such narratives. Especially journalists and educators ought to benefit from this work, since they often share a similar desire to catch the readers' interest, encourage the readers to explore the data by themselves and still learn the story intended by the author.

To summarize, our two primary contributions with this research are:

- Provide a tool for non-experts to access, explore and gain knowledge from The Swedish Migration Agency's public refugee data in Sweden
- Provide insights about reader-driven stories, narrated by non-experts, especially along the dimensions of *reader engagement* and *learning*

1.3 Research Question

As described above, in this thesis we set out with a desire to learn more about reader-driven stories and if *AsylKoll* can be used as tool for facilitating them. More specifically, we try answer the question:

How well does *AsylKoll* support the articulation of reader-driven stories, for non-experts, grounded on geographic-centric data?

The term *non-experts* is a loosely defined term used in this thesis to describe the general public, which has no particular experience or knowledge about data analysis, statistics and data visualization. Moreover, we target the Swedish population and therefore, the language in the user interface is Swedish.

By trying to answer the question posed above, we also hope to gain further insights about reader-driven stories and what non-experts need for deriving data-driven stories of their own.

1.4 Outline

This thesis is structured as follows. In the next section, Section 2, we will list and present the most relevant work done in the area of narrative visualization. Then, the tool is presented in greater detail, including the data which it visualizes, in Section 3. Following this, in Section 4, we will describe the method used to evaluate *AsylKoll*. In Section 5, we present the results of our user-studies. Subsequently, the results are analyzed and discussed in Section 6. Lastly, the thesis is summarized along with our conclusions and reflections on future work in Section 7.

2. RELATED WORK

Below we list and discuss previous work conducted in the same research area as this thesis and how they relate. We start by defining the scope of narrative visualization.

2.1 Narrative Visualization

Narrative visualization is a young and emerging field within information visualization. Due to this, the subject *narrative visualization* has been loosely defined so far, and still is. Bongshin Lee [2015] tries to give one definition of visual data stories as something that fulfills all of the following three conditions: i) a story that consists of story pieces backed up by data; ii) most of these pieces are visualized to support one or more intended messages; and iii) the pieces are presented in a meaningful order or with some connection between them to support the author's high-level communication goal. Thus, we can say that this definition emphasizes the importance of data, visualizations and presentation of them. We adhere to the definition of Bongshin Lee [2015] in this thesis, however, since we are focusing on reader-driven stories, *the author* and *the intended messages* will in most cases mean the readers and their intended messages. The only high-level communication goal that we, as authors have, is to spread a fact-based world view. In particular, to spread a fact-based view of the asylum status in Sweden, during 2015.

Looking back in time, one of the first to acknowledge the potential of storytelling in information visualization were Gershon and Page [2001]. In the paper they present a number of methods that can be used to build a story effectively using visualizations, which in their case is a story about a hostage situation. Even though the authors do not actually describe any data visualizations in detail, they argue that storytelling is a particularly useful tool for conveying key information to the reader. In the same way, we believe that

discovering those stories on your own can be a powerful method for readers to grasp and understand facts about the world.

Comparing to classical information visualization there is an important distinction between it and narrative visualization. The former tends to focus on analysis and the ability to explore the datasets which it maps, while the latter is more focused on the presentation and communication of insights concealed in the data. The implications of this has been that information visualization in general have focused on experts (e.g. statisticians, scientists), who might be used to working with data and need tools to explore it accurately. Narrative visualization is different, since it often targets non-expert readers that are not used to data analysis or statistics. Indeed, this is the audience we target in this study. This shift in target groups has important implications when constructing the stories. One important implication, recognized in several studies, being that of the text accompanying the visualizations [Borkin et al. 2016; Segel and Heer 2010].

In their paper, Segel and Heer [2010] review the design space of narrative visualization and classify different genres. The authors identify seven genres and give some general advice on how one can design narrative visualizations. In addition, they reflect on ways of balancing *author-driven* versus *reader-driven* stories. Segel and Heer [2010] note that while storytelling, historically, have had more of an author-driven structure, new information visualizations open up for interactivity and reader-driver stories. This is the area of research that we explore in this thesis.

The design space is, however, only one side of narrative visualization. Another equally important aspect is what effect the stories have on the readers. This is touched upon by Hullman and Diakopoulos [2011], when they investigate how different rhetorical decisions regarding how to frame a story, something the authors call *visualization rhetoric*, affect readers' interpretations of the story. This research has been extended by Hullman et al. [2013] into showing how to effectively sequence the visualizations in a story. They find that it is most effective to sequence by a temporal ordering and worse to sequence on granularity. Conclusions that we draw on when designing AsylKoll.

2.2 Reader Engagement

With the interactive information visualizations available today a common goal of designers is to create visualizations that encourage the reader to engage with the data and explore it. Something that is particularly essential if we want to create a tool which facilitates reader-driven stories. One example of increasing reader engagement is given in Diakopoulos et al. [2011], where the author tries to accomplish this through the usage of game mechanics. This approach has some interesting implications. By constructing the story as a game, the story is governed by the structure of the game, however, not bound by a linear sequencing as stories usually are. The game structure of the stories is found to increase the reader engagement, while decreasing the potential for insights and learning.

Another suggestion to increase reader engagement can be found in Yousuf and Conlan [2014]. The authors present their framework, VisEN, which dynamically creates personalized narratives as the reader explores the data. They find that these automatically generated and personalized stories have a positive impact on the reader's learning by increasing and improving their engagement with the data. Related to this research, albeit not exactly the same, is that of Boy et al. [2015]. Contrary to Yousuf and Conlan [2014], they find that the use of stories in relation to interactive visualizations do not increase the reader engagement.

Moere and Purchase [2011] argue for the importance of aesthetics when designing information visualizations. The authors mean that a *"highly aesthetic representation may compel the user to engage with the data"*. To assist in this work they provide a model of design in information visualization. The model defines three design domains - visualization practice, visualization studies and visualization exploration - which all have specific implications for design considerations. Moreover, aesthetics is an important factor to take into account when trying to create memorable visualizations.

All in all, it is safe to say that current research cannot really give a good answer on how to increase reader engagement. While the research described above in general focuses on the presentation of data and the design of the visualizations, we will in this thesis instead try to increase reader engagement by focusing on the data, i.e. by choosing geographic-centric data made relevant to the user.

2.3 Learning

Constructing stories that the reader learns from and remembers, is of interest for anyone trying to convey a message. There is an ongoing debate in the research society on how to design memorable visualizations. For a long time the consensus in information visualization was to keep the charts and visualizations as plain and minimalistic as possible. Implying that decorative and other non-related imagery, also called "chart junk", ought to be avoided [Tufte and Graves-Morris 1983]. Bateman et al. [2010] go against this consensus and argue that chart junk does not affect readers interpretation accuracy but it does, however, increase the recall of the visualizations.

Borkin et al. [2013] and Borkin et al. [2016] draw similar conclusions when investigating what makes visualizations memorable. They find that very plain charts and visualizations (compared to more uniquely designed) are difficult for readers to remember. In addition, they find strong evidence for the importance of using titles and annotations to make the story memorable. However, the authors also acknowledge the difficulty in ensuring that the readers remember the *right* thing: *"We do not want just any part of the visualization to stick (e.g. chart junk), but rather we want the most important and relevant aspects of the data or trend the author is trying to convey to stick."* [Borkin et al. 2013].

We will take a more conservative stand with regards to chart junk in this thesis, i.e. avoiding it and instead sticking to rather minimalistic charts, for reasons that will be explained in greater detail in 3. We are, however, guided by the research underlining the importance of the text accompanying the visualizations, especially the titles.

3. ASYLKOLL

In this section we present our tool, *AsylKoll*. A tool for exploring the datasets published, on a monthly basis, by The Swedish Migration Agency. AsylKoll is built targeting non-experts and the implications of this decision will be discussed later in this section. Moreover, as often is the case when designing information visualizations, a guiding mantra through the designing process has been that given by Shneiderman [1996] *"overview first, zoom and filter, then details-on-demand"*. However, before we get into the details of the visual mappings and the motivation behind them, in Section 3.2, we start by briefly presenting the data shown in the tool.

3.1 The Data

All the data which is visualized or presented in some way, through AsylKoll, comes from three providers: The Swedish Migration

Agency, Statistics Sweden (SCB) and Ipsos [Agency 2015; Sweden 2015; Ipsos 2015].

As explained in the introduction we chose to work with immigration data since it is such a heavily debated subject and rated as the most important political topic by Swedes in 2015 [Television 2016; Delin 2016]. We also decided to include population data from SCB to use in conjunction with the immigration data, to enable comparisons and put the immigration data in a local context. In addition, we incorporated some poll data on the public opinion regarding refugees, since we believed that adding Swedes perceptions on this subject would help give a better overview of the refugee status in Sweden.

All the data was cleaned and preprocessed using Python Pandas [Pandas 2016].

3.2 The Visuals

3.2.1 Technical Details. AsylKoll² is built on HTML, CSS and Javascript (jQuery, D3.js), with D3.js constituting the core of the visualizations [Bostock 2012]. The motivation behind making AsylKoll a web application is accessibility. Since the essential goal is to spread a fact-based world view, those facts should be distributed to an as large share of the population as possible. Also, it is well aligned with Gapminder's vision to be "*a modern 'museum' that helps making the world understandable, using the Internet*" [Gapminder 2016].

D3.js is well suited for web visualization, since it is a powerful library for manipulating DOM documents which, in addition, is supported natively by most modern browsers. Some visualizations in the tool are built directly using D3 while others, simpler visualizations, such as bar charts and columns charts, are created using C3.js [Tanaka 2014]. C3 is a visualization framework built on top of D3.js. The reason for using C3.js is its default support for filtering and details-on-demand, while being customizable and easy to set up and use.

Furthermore, accessibility is enhanced by the complete responsiveness of the site. All the charts, including the web page itself is responsive.

3.2.2 General Ideas. As an attempt to minimize the time spent on figuring out how the visualizations work and what they show, we have tried to limit the complexity of the visual mappings by using well-known types of charts. Such as maps, bar charts and column charts. This is a decision which most likely will decrease the memorability of the charts themselves, but as AsylKoll is a tool for the users to discover stories by themselves, we believe that an innovative design might accidentally take attention away from this task. Especially since it is hard to ensure that users remember the *right* thing [Borkin et al. 2013].

To further assist the user we have drawn on previous research and spent time on trying to use descriptive titles and keeping the explanatory text as precise and short as possible [Borkin et al. 2016, 2013; Segel and Heer 2010].

Learning from Hullman and Diakopoulos [2011] and Hullman et al. [2013], we have used temporal ordering as often as possible, to ease the cognitive processing of the data.

AsylKoll consists of two main views - *the arrivals view* and *the asylum view*. These are described one by one below. Also, as mentioned earlier in this section, the design process has been guided by the visualization mantra "*overview first, zoom and filter, then details-on-demand*" and as we describe the views we will

do so along the tasks discussed in Shneiderman [1996], namely: overview, zoom, filter, detail-on-demand, relate, history and extract. Note, however, that AsylKoll is in an early stage of the development phase and thus, the tool does not support all of these tasks.

3.2.3 The Arrivals View. The arrivals view shows data on asylum seekers who have been granted asylum and have been placed in some Swedish municipality. This group of asylum seekers is called *arrivals* in this thesis. The arrivals can end up in a municipality for five different reasons, defined by The Swedish Migration Agency. The reasons are briefly described below.

- (1) **Quota Refugee** - a refugee that has been elected abroad through a collaboration between The Swedish Migration Agency and the UN.
- (2) **Arranged Accommodation** - asylum seeker who has been granted resident permit and then been placed in a municipality by The Swedish Migration Agency or The Swedish Labor Agency.
- (3) **Self-Arranged Accommodation** - asylum seeker who has been granted resident permit and who has arranged with accommodation by himself or herself.
- (4) **Secondary Immigration** - a family member to an asylum seeker who has been granted resident permit and is already living in some municipality in Sweden.
- (5) **Miscellaneous** - mainly individuals who have been granted resident permit without applying for asylum.

The starting frame of the arrivals view shows an overview of where all the arrivals have ended up in Sweden, both in actual numbers and per capita (see Figure 1). To support the understanding of this data the tool makes use of a map of Sweden, next to a bar chart. The map is a choropleth, where the counties are colored by number of arrivals per capita. The bar chart accompanying the map shows the number of arrivals per county, in actual numbers. This initial frame provides an *overview* of the numbers of arrivals in Sweden during 2015 and the two graphs together enable the users to *relate* per capita numbers to absolute numbers.

The bar chart is a stacked bar chart with two options for grouping. One where the data is grouped by reasons for arrival (see list above) and one where the arrivals are grouped by age. Also, it is possible to *filter* the data in the bar chart by hovering and clicking the legend.

Moreover, the view provides a drill down feature, down to municipality level, by clicking on a specific county. Hence, it is possible to *zoom* and *filter* the data by county. When drilling down the map itself will zoom in on the county and color municipalities based on arrivals per capita instead (see Figure 2). The rest of the counties are colored gray in order not to steal focus from the selected county. The bar chart will also update. It shows the same data as earlier but per municipality instead, in the selected county.

In the zoomed in view, another column chart will appear too, which shows the arrivals by country of origin, in the selected county (see Figure 3). This will give a better *overview* of who has moved to that particular county and also humanize the arrivals data.

Finally, the map and the bar chart both give the user *details-on-demand* on mouse-over, in the zoomed in view as well as the zoomed out view.

3.2.4 The Asylum View. The other main view, the asylum view (see Figure 4), is made up of two charts and some key facts expressed in plain text. The main chart is a combination chart consisting of a column chart and a line chart. The columns express the

²<https://people.kth.se/~cjba/thesis/>

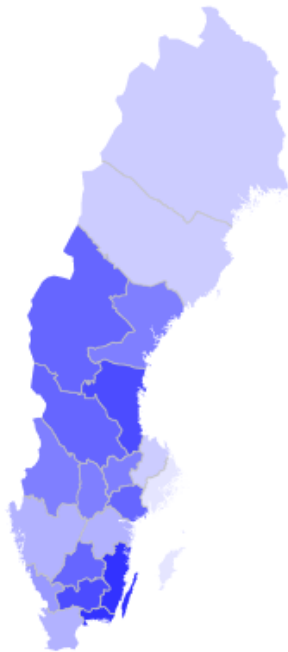
Vizylum

← Arrivals Asyl →

Arrival: Someone with a foreign citizenship who has been granted a residence permit in Sweden as a refugee, in need of protection, particularly distressing circumstances or as a relative (secondary immigration).

Arrivals per Capita

The map below shows Sweden's counties and municipalities. They are colored based on the number of arrivals per capita they have received. Darker color means a higher number of arrivals per capita. Click on the map to get statistics on municipality level.



Arrivals in Absolute Numbers

The Swedish Migration Agency classifies the arrivals in five different categories. These are based on the reason for a certain refugee to arrive in a certain municipality. The bar chart below shows the total number of arrivals, ordered by geographical area and category.

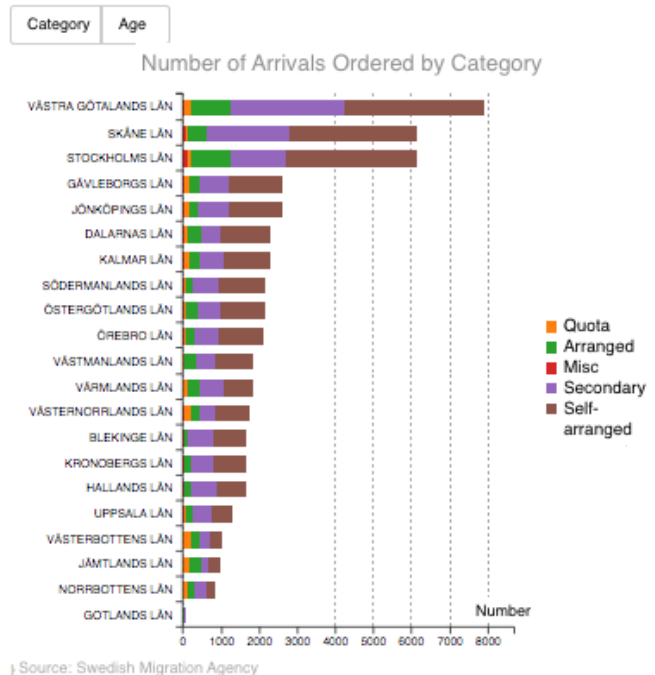


Fig. 1. Starting frame of the arrivals view. Showing a choropleth map, colored by arrivals per capita, and a bar chart showing the number of arrivals in absolute numbers. Note: the text has been translated to English in this thesis, however, the web app is in Swedish. For a short video showing the tool, follow this link: <https://vimeo.com/187571807>.

number of asylum applications, the number of decision made in asylum cases (on whether or not to grant asylum) and the number of people who was granted asylum. The data is presented by month, to show the *history* of the asylum immigration over the year. The line chart which accompanies the column chart shows the average handling time per case, in days. Together this combination chart gives a thorough *overview* of the status of the asylum immigration in Sweden during 2015. It provides insights about when most asylum seekers arrived to Sweden, the efficiency of The Swedish Migration Agency, the number of asylum seekers still waiting for an answer and more.

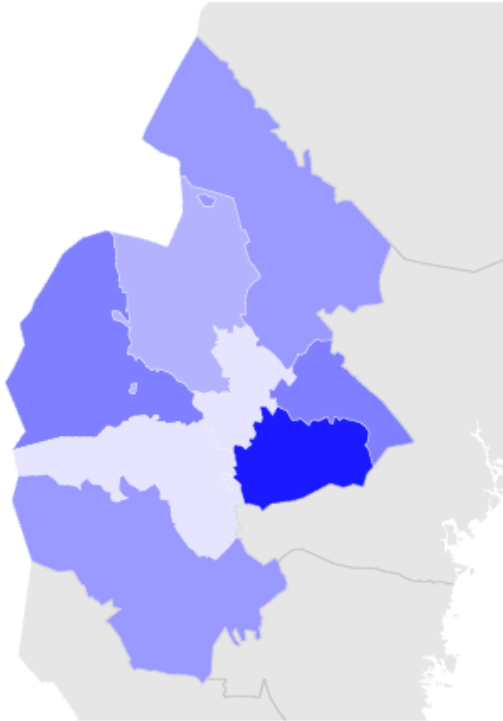
Alongside the main chart described above is another column chart, which shows the Swedish opinion regarding whether or not to accept more refugees in Sweden. The same poll was conducted at three different occasions during 2015 and put together they show a short *history* of how the opinion varied in Sweden during the

year. Since it is presented together with the asylum chart it is possible for the user to *relate* the asylum situation in Sweden with how the Swedes perceive the development. In addition, both of the charts discussed above have options for *filtering* the data and getting *details-on-demand* by doing a mouse-over.

Furthermore, the asylum view also has some key figures presented as plain text. The figures are highlighted with a bigger font and a different color. These key facts yield a simple, but powerful, way to give the user an *overview* of the migration data and it is basically a way of letting the user *extract* key facts from the data. It is powerful in the sense that it offers several different insights without requiring any interpretation of visual mappings and thus, is cognitively cheap.

Arrivals per Capita

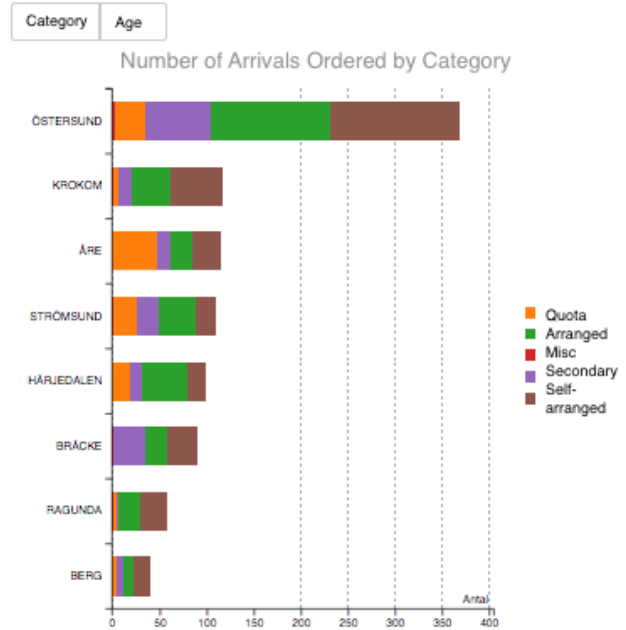
The map below shows Sweden's counties and municipalities. They are colored based on the number of arrivals per capita they have received. Darker color means a higher number of arrivals per capita. Click on the map to get statistics on municipality level.



Source: Swedish Migration Agency, Statistics Sweden

Arrivals in Absolute Numbers

The Swedish Migration Agency classifies the arrivals in five different categories. These are based on the reason for a certain refugee to arrive in a certain municipality. The bar chart below shows the total number of arrivals, ordered by geographical area and category.



Source: Swedish Migration Agency

Fig. 2. The arrivals view when zoomed in on a particular county. The municipalities within the selected county is colored by arrivals per capita. The bar chart is also updated, showing data on municipality level instead. Note: the text has been translated to English in this thesis, however, the web app is in Swedish. For a short video showing the tool, follow this link: <https://vimeo.com/187571807>.

4. METHOD

In this section we will explain the methodology that we decided on, to research the question posed in Section 1.3. First, we can note that there are several ways of testing user interfaces. Choosing a methodology is not dependent only on the research question and what you want to test, but also on the interface itself, in particular where in the development phase you are. AsylKoll is in a very early stage of development, which needs to be taken into account when running the user-studies. If there are many bugs and the design is too rough, there is a risk that these obstacles disturb the users to such an extent that they are not able to properly test the interface. Thus, producing insufficient amount of data with such a quality that it is valuable to analyze.

One way of getting around this problem of early stage testing, is presented by Kinnaird and Romero [2010]. A method used in the development of *Connect 2 Congress* [Kinnaird et al. 2010]. The authors conducted their experiment by using focus groups, but instead of letting the users interact directly with the visualizations, two *facilitators* participated in the study as well. The facilitators were familiar with the software and operated it for the users, in addition to recording the session and taking notes.

However, AsylKoll is built as a web interface and intended to be a window into Swedish immigration data, for non-experts. Thus,

one key aspect is to analyze the behavior of first time users, who for some reason enter the website. Therefore, using a facilitator to guide the participants of the user-studies, might take away this effort to mimic the behavior of new users. Instead we have tried to develop and test the user interface, to such an extent that, it is simple and fast enough. Meaning that any imperfections in the software are so small that they will not interfere with the evaluations of the user-studies. To help understand how the users think when interacting with the software, we use a method called the *Think Aloud Method*, which we describe in the next section [Somerén et al. 1994].

4.1 The Think Aloud Method

Solving a problem can be thought of as the process of finding out the answer to a question without knowing the answer in advance. This is, in essence, what we are asking users to do when we want them to find their own stories, i.e. create reader-driven stories. With the help of a tool, in this study AsylKoll, we want the users to articulate stories of their own, without prior knowledge of what those stories might be, in order to make them learn and get a more fact-based world view. Therefore, for us to understand what users need to facilitate this learning process, we need to know what their *thought processes* look like. Because given a task or a problem, people can end up with the exact same answer but the thought pro-

From Where Did They Come?

Column chart showing the number of arrivals in the selected country, ordered by country of origin. Logarithmic scale.

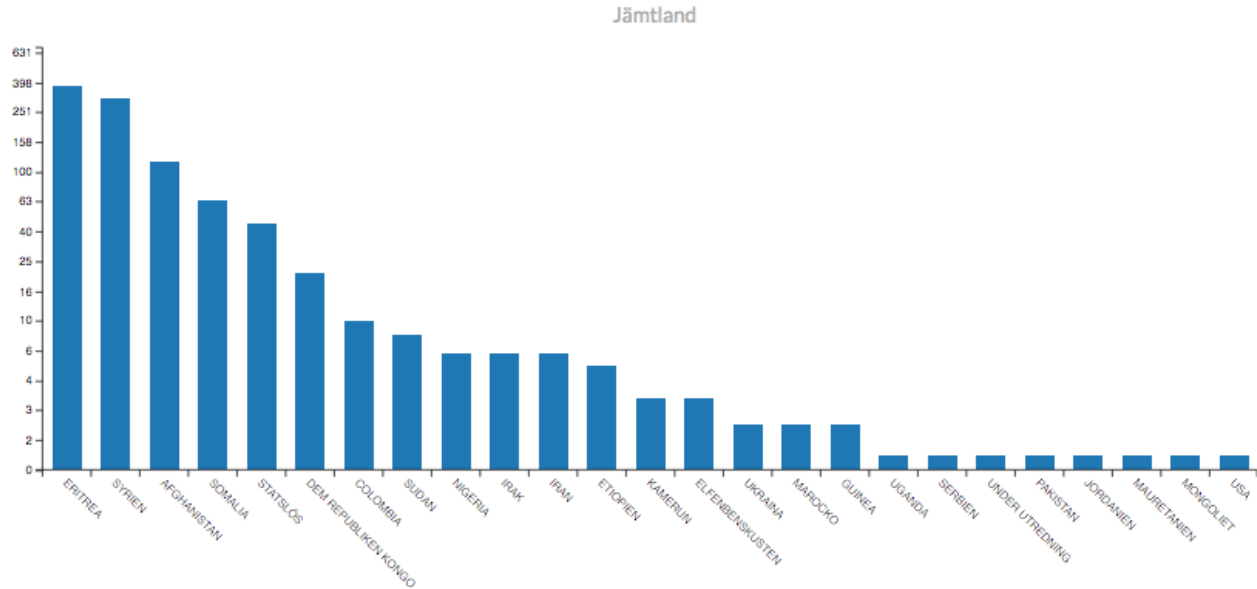


Fig. 3. The column chart that appears when zooming in on a county. The chart shows the number of arrivals per country of origin, sorted in descending order. Note: the text has been translated to English in this thesis, however, the web app is in Swedish. For a short video showing the tool, follow this link: <https://vimeo.com/187571807>.

cess of how they arrived at that answer can look very different. The Think Aloud Method is well suited for making it possible to evaluate these cognitive processes and the differences of them, between individuals.

As the name suggests, the method works such that the participants are asked to continuously think aloud and speak their mind, while using the software. The participants are asked to describe how they perceive the visualizations; how they go about understanding them; what is surprising about both the data and the visualizations; what they think is missing; what additional data they would like to compare the existing data with and so forth.

The data captured from the think aloud experiments is called *verbal protocols*, which simply is some kind of record of what the participants are saying. This data can be captured in various ways, e.g. taking notes, audio or video recording. The verbal protocols are later coded according to a coding scheme, defined to help evaluate parameters of interest. The coding scheme used in this thesis is described in Section 4.2.2.

4.2 Experimental Setup

4.2.1 User-Studies. There are many ways of conducting a user-study with the Think Aloud Method. One common way is to evaluate single users, while they interact with the software. However, it can be difficult for people doing this type of experiment for the first time to be comfortable and speak their mind freely. For this reason we decided to conduct the study with groups instead and use a variant of the Think Aloud Method, called *Dialog Observation*.

In Dialog Observation the participants solve tasks together. Tasks where they need to collaborate and communicate in order to solve them. This method was chosen since it is generally simpler for people that are not used to doing individual think aloud sessions. The

discussions help the participants communicate in a more relaxed way, thus, providing more data for evaluation.

Two user-studies were performed. The first with four participants and the second with three. In both user-studies one administrator was also present, taking notes and recording the session on video. The administrator did not participate in the discussion, only reminding the participants to keep on talking if they went silent for too long. The video recordings were later used for analysis.

During both user-studies the participants were given the same two tasks:

- (1) Explore and familiarize yourself with the data and the visualizations
- (2) Construct four tweets that you find interesting and newsworthy

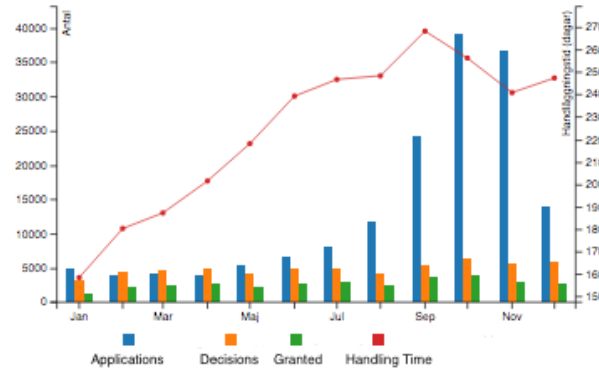
The participants were not given the second task until the first one was finished. There were no strict time limits for any of the tasks and the first task was ended by the administrator, when the participants seemed to have explored the data sufficiently and the discussion naturally came to an end. Then they were given the second task, with the requirements that the tweets had to be fact-based and that every tweet had to include a screenshot from the tool, supporting the text in the tweet. All tweets can be found in appendix A.

Someren et al. [1994] lists two important conditions to consider when deciding on which tasks participants are asked to perform:

- Task should be at a level of difficulty that is appropriate for the subjects with respect to the cognitive process
- The task should be representative with respect to the cognitive process involved

Asylum Applications and Handling Time

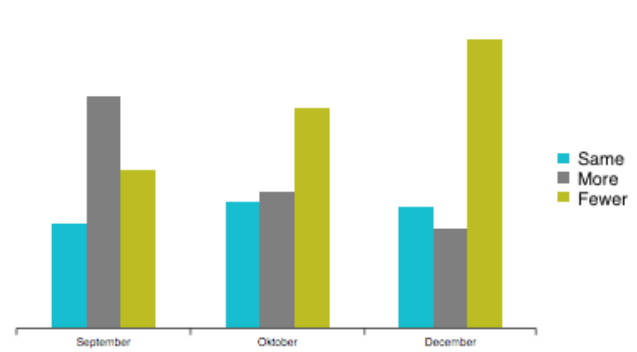
The figure below shows number of applications, decisions and granted applications, together with the average handling time per application.



Source: Swedish Migration Agency

Swedish Opinion

In a poll, performed three times during 2015, Swedes were posed with the question "Should Sweden accept more, fewer or the same amount or refugees?" The result of every poll is shown below.



Source: DN/Ipsos

2015 in Figures

162877

applied for asylum in Sweden.

48149

were women.

35369

were unaccompanied children.

32631

were granted asylum in Sweden.

Source: Swedish Migration Agency

55%

of all applications were granted.

229

hours was the average handling time.

Syria, Afghanistan and Iraq

were the top three citizenships among the asylum seekers.

Syria, Eritrea and Stateless

were the top three citizenships among those who were granted asylum.

Fig. 4. The asylum view. The visualization on the left shows a summary of applications, decisions, granted applications and average handling time during 2015. On the right there is a chart showing the Swedish opinion regarding refugees and at the bottom some key figures are presented in text. Note: the text has been translated to English in this thesis, however, the web app is in Swedish. For a short video showing the tool, follow this link: <https://vimeo.com/187571807>.

The cognitive process here can be understood as the process of constructing stories of your own, through data exploration. We would argue that the two tasks, presented earlier in this section, meets the conditions listed above. The first task is a prerequisite for understanding how a new user would perceive and interact with the tool, when presented with it for the first time. The second task is representative to the process of creating stories, since every tweet is a short story of its own. It is, however, hard to assess the correct level of difficulty. Nonetheless, writing a tweet is better than instead of e.g., a short news article, which would be a possible option. Because a tweet is so short that it ought not to be an obstacle for participants who are not used to or comfortable with writing texts. At the same time, being limited to only 140 signs forces the participants to carefully construct the sentences, in order for them to be descriptive and newsworthy. Requiring the participants to always include a picture makes it easier to see how they use the tool and what they find interesting in the data.

Finally, given the fact that the data concerns a very polarized political topic in Sweden and that the participants were asked to

agree on the tweets, they had to collaborate and discuss to complete the task.

4.2.2 Coding Scheme. To be able to evaluate the verbal protocols, i.e. the video recordings, they were *segmented* and later *classified* [Somerén et al. 1994]. The segments were delimited by either a natural end to the discussion, i.e. longer pause or a change of topic. Hence, a segment can be defined as a coherent discussion about one topic.

Next, the segments were classified, according to the coding scheme shown in Table I. The segments were classified in two levels. A high-level, which classifies *what* was being discussed (VIS, DA and ME), and a low-level, which classifies *how* the discussed feature/subject affected the participants (RE, IN and ME). This is a simple coding scheme, focused on evaluating the parameters presented in the research question (see Section 1.3), and not intended to be exhaustive.

4.2.3 Survey. In order to capture whether the participants learned anything from the tool, how well it facilitated data exploration by non-experts and if they found the data interesting, the

Table I. Coding scheme used when analyzing the verbal protocols.

Code	Name	Description
VIS	Visualizations	Discussion about the visualizations
DA	Data	Discussion about the data
ME	Meta	Everything that is not classified as DA, VIS, RE or IN
RE	Reader Engagement	Discussions/comments that say something about reader engagement
IN	Insights	Discussion/comments about how the participants gain insight/learn from the tool

participants were asked to answer a short survey at the end of the session. The survey consisted of four Likert scale questions (scale 1-5). The questions in the survey were:

- (1) Did you learn something new about the asylum immigration in Sweden?
- (2) Did you find anything in the data surprising?
- (3) Was it difficult to understand the visualizations?
- (4) Was it difficult to interact with the tool?

The result of the survey is shown in Section 5.2.

5. RESULTS

In this section we present the results of our two user-studies, including the surveys. The results are discussed in the next section, in Section 6.

5.1 User-Studies

In Figure 5 we present the statistics from the user-study. More specifically, we show the distribution of the high-level codes, derived during the analysis of the segments. The distribution of high-level codes are also shown in Table II and low-level codes in Table III.

All tweets produced during user-study 1 and 2 can be found in appendix A.

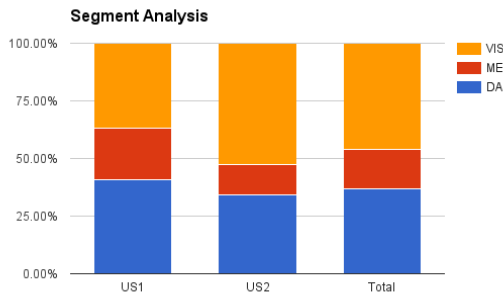


Fig. 5. Distribution of high-level codes from both user-studies (VIS - discussion about the visualizations, DA - discussion about the data, ME - everything that is not classified as DA or VIS).

5.2 Survey

The results of the survey that the participants were asked to complete at the end of the session, are shown in Figures 6 - 9 below.

Table II. High-level codes, in both user-studies.

Code	US1	US2
VIS	10	20
DA	11	13
ME	6	5

Table III. Low-level codes, in both user-studies.

Code	US1	US2
ME	6	5
RE	4	5
IN	17	28

Do you feel that you learned anything new about the Swedish reception of asylum seekers? (7 responses)

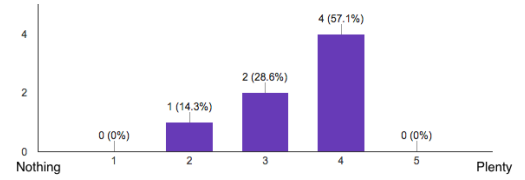


Fig. 6. The results from question 'Do you feel that you learned anything new about the Swedish reception of asylum seekers?'. Scale 1-5, where 1 is 'nothing'.

Was there anything in the data that surprised you? (7 responses)

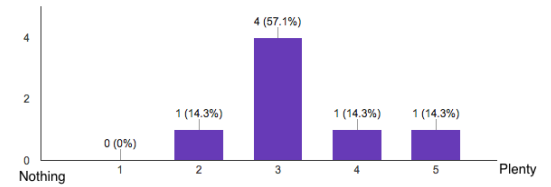


Fig. 7. The results from question 'Was there anything in the data that surprised you?'. Scale 1-5, where 1 is 'nothing' and 5 is 'plenty'.

Did you find the visualizations difficult or easy to interpret? (7 responses)

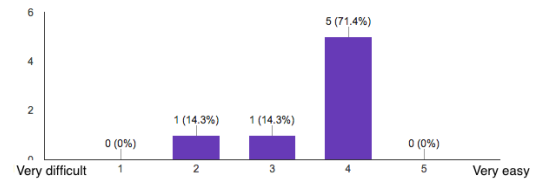


Fig. 8. The results from question 'Did you find the visualizations difficult or easy to interpret?'. Scale 1-5, where 1 is 'very difficult' and 5 is 'very easy'.

6. DISCUSSION

In this section we will discuss and reflect on the outcomes of the analysis of the data that was gathered from the two user-studies and surveys. The general impression is indeed that AsylKoll al-

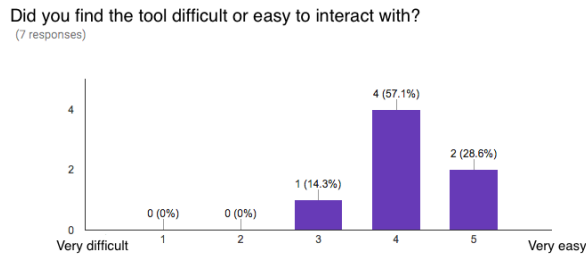


Fig. 9. The results from question 'Did you find the tool difficult or easy to interact with?'. Scale 1-5, where 1 is 'very difficult' and 5 is 'very easy'.

ready, at its current state, is able to help users articulate their own data-driven stories. During the user-studies the data and the visualizations spurred the participants into several different discussions about asylum immigration in Sweden. About cause and effect and how to utilize the data to create better integration. All in all, it was obvious that the topic was one of great interest.

Breaking down the discussions into segments and coding these yielded the result shown in Figure 5. From that figure we see that even though the visualizations were discussed the most, there were almost as much discussions about the data itself. That is a positive result, since the data is key to a fact-based world view. The visualizations are only ways to facilitate the process of spreading that world view and the optimal result would be that all discussions were about the data. However, the result shown in Figure 5 is especially good, given that the participants were in fact encouraged to discuss the visualizations, to provide data about improvements of the tool. Nevertheless, the participants quickly got a grasp of how to use the tool and the data. Something that is supported by their answers to question 3 and 4 in the survey (see Figure 8 and Figure 9).

Furthermore, an apparent finding, when analyzing the users' behavior, is the importance and the difficulty of using explanatory text together with the visualizations. A finding which partly confirms what has earlier been observed in Borkin et al. [2016]; Segel and Heer [2010]. In both focus groups, the importance of text was shown through the fact that reading the text was the first action for all participants. However, the users read the text quickly and not very carefully, which resulted in them missing several important instructions. Even though the explanatory text in the tool was only a few sentences in total. Hence, the text did little to help the participants figure out how to use the visualizations, contrary to what much of the previous research say. We see this as an apparent difficulty of using text.

Therefore, we would argue that one should **use text as little as possible** by trying to create the visualizations where very few text-based instructions are needed. Show instead the interactivity of a visualization by using various effects, e.g. change color or size on mouse-over, or use an introductory video that shows the user how the visualizations work. If text is needed, emphasize important instructions somehow, e.g. size, color or font style.

When looking at the stories produced during the two user-studies, i.e. the tweets (see appendix A), there are several interesting observations we can make. First, we note that the tweets from the two user-studies are similar, even though the studies were performed independent from each other. Figure 12 and Figure 16 show that **finding extremes** and **comparing** are important features when finding stories. Next, from Figure 13 and Figure 14, we see that both groups found the data on the arrivals' country of origin

interesting. In general, the participants seemed to be interested in finding stories *about* those who came to Sweden during 2015. E.g. besides country of origin, the participants liked that age data was provided too. We interpret this as an interest to **humanize the data**. Indeed, if the data is in fact human data, as in AsylKoll, interest can be evoked by attaching human attributes to the abstract statistics and visualization.

There were also some indicators missing in AsylKoll. One thing, which was mentioned on different occasions by both groups, was the interest to correlate the data with media somehow. Since the topic of asylum immigration has been a hot topic in Sweden during the last year, media has written a lot about refugees and everything related to refugees, e.g. conflict zones. For this reason, it would have been good to include some kind of data on relevant news articles. Perhaps a window similar to what was included in Connect 2 Congress [Kinnaird et al. 2010]. Another indicator that both groups asked for and discussed was income. In fact, **income seems to be an indicator which attracts a lot of interest**.

A notable obstacle for having users derive reader-driven stories, related to media and the use of statistics in general, surfaced during the user-studies. Multiple participants tried to figure out what the author was trying to say with the data and the visualizations, despite that the tool was built to present the data as plain as possible, in order to avoid the presence of any author-driven stories. This source criticism is a sound approach to statistics, since there are many abusive ways of using it. However, if the goal is to let users find their own stories, this attitude can take precious focus from the data and instead change it to finding stories that are not there. In worse case, the user might refuse to see the data as facts.

6.1 Reader Engagement

As has been mentioned earlier, in Section 2.2, in this thesis we wanted to see if we could increase reader engagement by focusing on the data, i.e. by using geographic-centric data relevant to the user. We would argue that this is indeed possible. The choropleth map in AsylKoll was the visualization that got the most attention. After reading text, which was the first action of all the participants, they immediately started to analyze the map. Discussing which county welcomed the most refugees and which who welcomed the least. The map also seemed to be easy for everyone to understand and extract information from.

Once the participants figured out that it was possible to zoom in on a county by clicking the map they wanted to explore the extremes and the county where they lived. Using local data did increase reader engagement and it seems like it might be a good idea to use as local data as possible. AsylKoll presents data on municipality level but in both focus groups the users wanted to explore on an even more local level. Thus, **making the data as user-centric as possible**.

There are also some conclusions which can be drawn regarding how to make users interact with the visualizations. In both focus groups the number of interactions were quite few in the beginning but as soon as they pass a certain threshold, the participants start clicking and exploring the visualizations much more. Therefore, it is important to ensure that the users are encouraged to interact with the visualizations as quickly as possible. E.g. if the users are presented with several different visualizations, it is advisable to let the users face simple, interactive visualizations at first. This will help the users if they are faced with a more complex visualization later, since they will be more willing to explore it.

A final observation was the positive correlation between visualizations with **details-on-demand** and how the participants engaged with the data and the visualizations, i.e. the reader engagement.

6.2 Learning

By looking at Figure 6 and Figure 7 we can conclude that the participants felt they did learn new things about asylum immigration in Sweden, by using AsylKoll. When analyzing the video recordings there are some features that are recurring for both groups, regarding what helps them gain insights and learn from the data and the visualizations.

One frequent request for example, mentioned earlier in this section, was the ability to **make comparisons** both within and between geographical-levels. E.g. comparing counties, comparing a municipality with a county, comparing a municipality with the country as a whole. Moreover, the participants were interested in **finding the extremes**, minimum and maximum, for all geographical levels. In addition, the provided **details-on-demand** had a positive impact on how the participants drew insights and learned from the tool. In addition, there was a need expressed to always present absolute numbers together with relative numbers. Absolute numbers alone evoke interest but **absolute numbers in conjunction with relative numbers provide understanding**.

7. CONCLUSIONS AND FUTURE WORK

In this thesis we set out to explore the domains of reader-driven stories by performing user-studies on a very basic tool, AsylKoll, which shows immigration data in Sweden. From this initial and limited study we have seen that even a simple and minimalistic visualization tool, such as AsylKoll, has the potential to spark fruitful discussions and raise questions, which fuel a need for further exploration. Essential requirements for having users successfully discover their own data-driven stories.

From our user-studies we have been able to gain both general insights about visualizations that support the articulation of reader-driven stories. However, we chose to focus on exploring how to create visualizations that maximize reader engagement and ensure learning. We found that reader engagement can be increased by doing the following:

- Center the data around the users
- Encourage users to start interacting with the visualizations as soon as possible to maximize the number of interactions
- Use a lot of transitions, mouse-over effects and details-on-demand

Similarly, we find that it is possible to boost learning by providing features which enable users to easily:

- Make comparisons
- Find extreme values
- Extract a lot of additional data through details-on-demand
- Explore absolute numbers in conjunction with relative numbers

However, this study is merely a first glance into the vast research space of reader-driven stories. Looking ahead, gathering quantitative data, instead of qualitative as in this study, will be essential. E.g. by publishing the tool on a web page with a lot of visitors and analyzing how users interact with the tool. The data generated would be valuable for several reasons. It would i) mimic first user behavior better than those group discussions used in this thesis; ii) possible impact that the administrator had on the participants would

be controlled; and iii) the data would come from a much wider sample of user types. Furthermore, a more advanced coding scheme, than the one used in this thesis, could be used by including more categories and, thus, providing more detailed insights about users' behavior.

In conclusion, in a world with an ever increasing supply of data and a continuous abuse of statistics and facts, we firmly believe that reader-driven stories have an important role to play in the quest of spreading a fact-based world view.

ACKNOWLEDGMENTS

We are grateful to the following people for resources, discussions and suggestions.

Mario; for support, coaching and positive energy, without which this thesis would have been a far greater struggle.

Ola; for the opportunity to work at such a great foundation. Angie, Fernanda, Gabriela, Jasper, Mattias, Max and Åsa; for a warm welcome, valuable insights and many laughs.

Elisabeth, Eric, Erik, Filip, Mariam, Mathias and Viktor; for lending the project your amazing brains and thus, providing the means needed to finish this thesis.

REFERENCES

- The Swedish Migration Agency. 2015. Current Statistics. (2015). <http://www.migrationsverket.se/Om-Migrationsverket/Statistik.html> Accessed: 2016-06-09.
- Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2573–2582.
- Petra Isenberg Sheelagh Carpendale Bongshin Lee, Nathalie Henry Riche. 2015. More than Telling a Story: A Closer Look at the Process of Transforming Data into Visually Shared Stories. *Computer Graphics and Applications, IEEE* 35, 5 (2015), 84–90.
- Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2016. Beyond Memorability: Visualization Recognition and Recall. *Visualization and Computer Graphics, IEEE Transactions on* 22, 1 (2016), 519–528.
- Michelle A Borkin, Azalea A Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Alfonso Oliva, and Hanspeter Pfister. 2013. What makes a visualization memorable? *Visualization and Computer Graphics, IEEE Transactions on* 19, 12 (2013), 2306–2315.
- Mike Bostock. 2012. D3.js. (2012). <https://d3js.org/> Accessed: 2016-06-09.
- Jeremy Boy, Françoise Detienne, and Jean-Daniel Fekete. 2015. Can Initial Narrative Visualization Techniques and Storytelling help Engage Online-Users with Exploratory Information Visualizations? (2015).
- Mikael Delin. 2016. Invandring den viktigaste frågan för väljarna. (2016). <http://www.dn.se/nyheter/sverige/invandring-den-viktigaste-fragan-for-valjarna-2/> Accessed: 2016-07-02.
- Nicholas Diakopoulos, Funda Kivran-Swaine, and Mor Naaman. 2011. Playable data: characterizing the design space of game-y infographics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1717–1726.

- Ryan Eccles, Thomas Kapler, Robert Harper, and William Wright. 2008. Stories in geotime. *Information Visualization* 7, 1 (2008), 3–17.
- Gapminder Foundation. 2015. Vizabi. (2015). <http://vizabi.org/> Accessed: 2016-06-09.
- Gapminder. 2016. About Gapminder. (2016). <http://www.gapminder.org/about-gapminder/> Accessed: 2016-06-09.
- Nahum Gershon and Ward Page. 2001. What storytelling can do for information visualization. *Commun. ACM* 44, 8 (2001), 31–37.
- Jessica Hullman and Nicholas Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on* 17, 12 (2011), 2231–2240.
- Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Daniel Fisher, and Eytan Adar. 2013. A deeper understanding of sequence in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on* 19, 12 (2013), 2406–2415.
- Ipsos. 2015. Swedish Public Opinion. (2015). <http://www.ipsos.se/> Accessed: 2016-06-09.
- Peter Kinnaird and Mario Romero. 2010. Focus Groups for Functional InfoVis Prototype Evaluation: A Case Study. *Beyond Times and Errors: Novel Evaluation Methods for Information Visualization, BELIV* (2010).
- Peter Kinnaird, Mario Romero, and Gregory Abowd. 2010. Connect 2 congress: visual analytics for civic oversight. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2853–2862.
- Robert Kosara and Jock Mackinlay. 2013. Storytelling: The next step for visualization. *Computer* 5 (2013), 44–50.
- Kwan-Liu Ma, Isaac Liao, Jennifer Frazier, Helwig Hauser, and H-N Kostis. 2012. Scientific storytelling using visualization. *Computer Graphics and Applications, IEEE* 32, 1 (2012), 12–19.
- Andrew Vande Moere and Helen Purchase. 2011. On the role of design in information visualization. *Information Visualization* 10, 4 (2011), 356–371.
- Python Pandas. 2016. Python Data Analysis Library. (2016). <http://pandas.pydata.org/> Accessed: 2016-06-09.
- Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *Visualization and Computer Graphics, IEEE Transactions on* 16, 6 (2010), 1139–1148.
- Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 336–343.
- MW van Someren, Yvonne F Barnard, Jacobijn AC Sandberg, and others. 1994. *The think aloud method: a practical approach to modelling cognitive processes*. Academic Press.
- Statistics Sweden. 2015. Statistical Database. (2015). <http://www.statistikdatabasen.scb.se/pxweb/en/ssd/?rxid=c546a375-ae6b-4a44-bd97-3e6d2f7f8d7e> Accessed: 2016-06-09.
- Masayuki Tanaka. 2014. C3.js. (2014). <https://c3js.org/> Accessed: 2016-06-09.
- Swedish Television. 2016. SVT News. (2016). <http://www.svt.se/nyheter/lokalt/vast/majoritet-ser-immigration-som-det-viktigaste-samhallsproblemet> Accessed: 2016-06-09.
- Edward R Tufte and PR Graves-Morris. 1983. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT.
- Bilal Yousuf and Owen Conlan. 2014. Constructing Narrative Visualizations as a Means of Increasing Learner Engagement.. In *HT (Doctoral Consortium/Late-breaking Results/Workshops)*.

APPENDIX

A. TWEETS

This section includes all the tweets produced during both user-studies.

Två flugor i en smäll. Anställ handläggare och lös arbetslösheten.

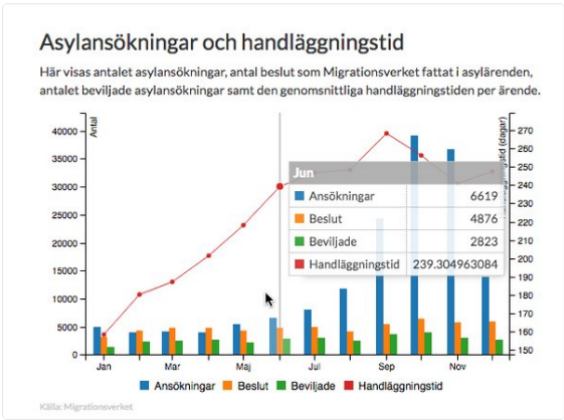


Fig. 10. Tweet from user-study 1.

Hur känns det här, Jimmy? #JimmieÅkesson

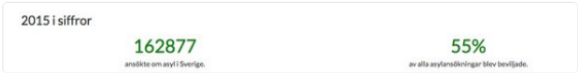


Fig. 11. Tweet from user-study 1.

Lika barn leka bäst!

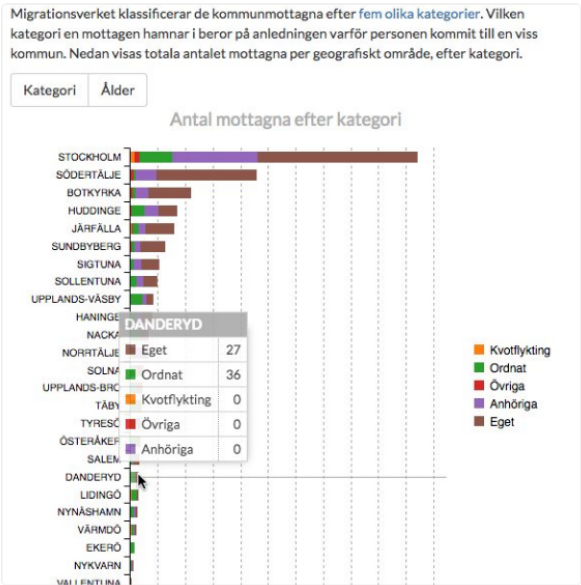


Fig. 12. Tweet from user-study 1.

Det ökande antalet asylsökningar är ej lika hårt kopplat till Syrien-krisen som media framställer det. #asylkoll

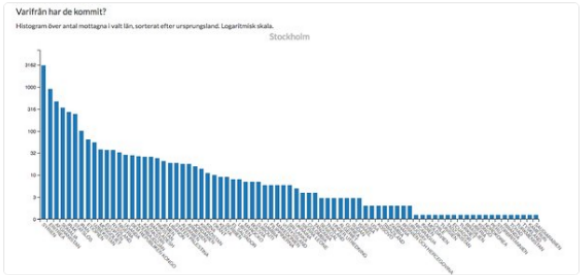


Fig. 13. Tweet from user-study 1.

Glöm inte att det pågår konflikter i andra delar av världen än Syrien.

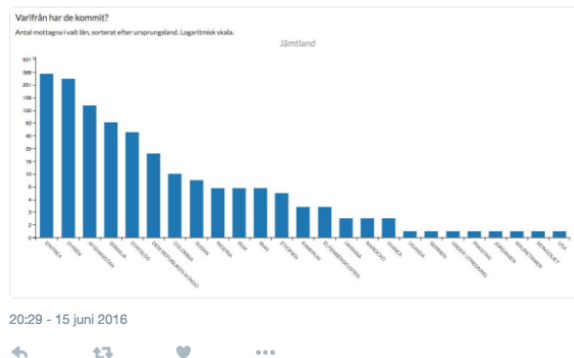


Fig. 14. Tweet from user-study 2.

Vaxholm - SÄMST i länet!

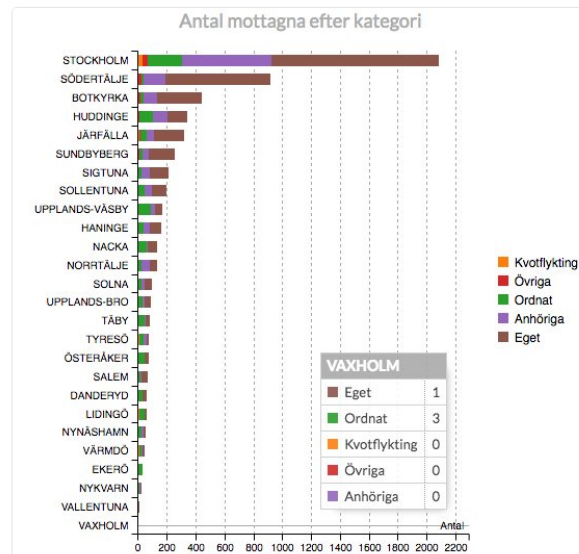


Fig. 16. Tweet from user-study 2.

Har du ändrat din åsikt?

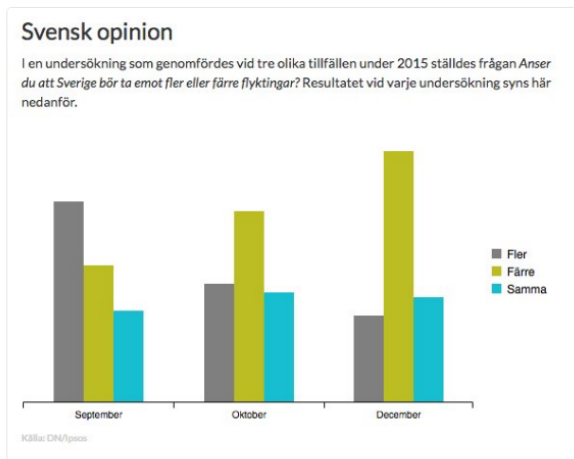


Fig. 15. Tweet from user-study 2.

Lägg inte pengar på gränskontroller, utöka Migrationsverkets resurser istället.

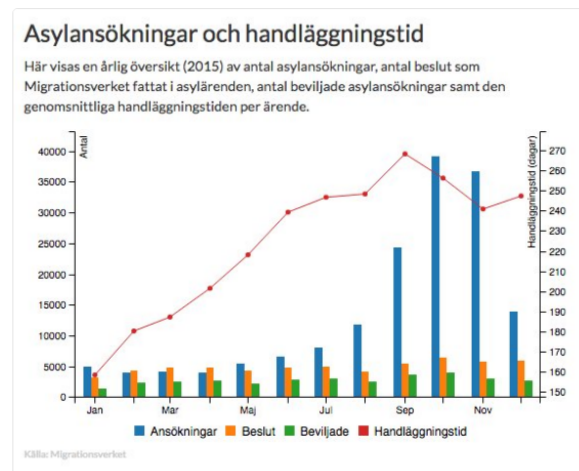


Fig. 17. Tweet from user-study 2.