IJPIS

EYES WIDE SHUT

- OR USING EYE TRACKING TECHNIQUE TO TEST A WEB SITE

ANNEGRETE WULFF
Statistics Denmark
2100 Copenhagen, Denmark
awu@dst.dk

Abstract

This presentation provides an introduction to the technique called 'Eye Tracking' which Statistics Denmark has recently used to test the usability of its website user interface. Eye tracking is a tool used to analyse "human - computer" interaction. The user's eye movements and fixation time are both registered when this technique is used. However, whether the person actually perceives what the eyes are fixed on is not documented. Statistics Denmark has tested this method on ten users who were asked to carry out particular tasks on the website. It is open to discussion whether the method is strong enough to be used independently. Some of the results were quite persuasive and supported the comments from other users. However, we suggest that eye tracking should not be the only method employed, but it should be used in combination with more traditional usability methods such as, "think aloud" and follow-up interviews.

Keywords: Usability, website, eye tracking, human computer interaction

1. Background

Statistics Denmark has for some years been considering the publication of statistics on the web prior to a paper version. The move from paper products to the Internet will consequently place the users in a new situation: where once only rather limited and previously processed data existed in paper format they now find themselves in the middle of the overwhelming information society.

In this "jungle" the users are left to their own devices to search, find, select, retrieve, comprehend and evaluate the information. To facilitate such a process it is of importance to create a logical information structure and use understandable terminology. However, what is logical and understandable for one user is not necessarily the same for someone else. The idea is, nevertheless, to build on some commonly used standard.

Thus, if it is not possible to deliver the best solution for everybody, the most commonly used standard solutions can be utilised in our web site. This, we believe, will make the user confident and guide him or her through the site.

The intention of our experiment was to follow up on this idea: do the users more easily find functions and information placed in a similar manner to many other sites? Will more help information on the page improve the understanding of the structure? Which parts of the site are most likely to be read?

The experience from earlier tests showed that users found it difficult to find the information. They required documentation and asked for download possibilities in Excel format. In our view navigation and functionality supporting these needs had already been set up [Wulff, 2006]. However, this was apparently not obvious to the

users. Other usability tests carried out by Statistics Denmark indicated that even a well agreed standard such as “bread crumbs” which showed where the user was and how to proceed in a backwards stepwise manner, was not perceived by any of those tested [Jensen et al., 2002]. This was one reason for agreeing to test the eye tracking methods on the web site. It was assumed that this could shed light on how users watch elements on our site.

2. How the Technique Works

In our test the eye tracker used is a Tobii 1750 connected to a Dell Precision 370 equipped with a 3.2 Ghz Intel Pentium CPU and 2 Gb Ram. It was built into a 17 " flat screen with a high quality web camera and light sources of “Near infra-red light”. The infra red light is reflected in the eye’s cornea and pupil. The combination of a reflection in the dark pupil with the reflection in the cornea is used to calculate the focus of the test user’s eyes. It is possible to eliminate the movements of the head.
Before the system is started it is calibrated to match each test person’s eyes.

The results are shown in “real-time videos” or “gaze replays”. They consist of recorded gaze plots that display the fixations and scan paths. A scan path can be understood as a “view route”. Another way to show the results involves “hot spots” which give a total view of the site. The red parts are those most looked at, while blue are only glanced at. (The scale goes from red, yellow, green to blue).

The test was carried out over three days for Statistics Denmark by The Royal School of Library and Information Science as a research project [Lund, 2006].

2.1. Test Persons

The people tested were 10 library students – 4 men and 6 women – with only very limited experience in the use of Statistics Denmark’s website. They were asked to find some information and carry out specific, defined tasks on Statistics Denmark’s web site.

A sheet describing the tasks was delivered to the students, one at a time. After having completed the task they were asked to mark and comment upon their navigation in the correct order as accurately as possible. This was completed using paper with downloads from the web site. This method, Retrospective Think Aloud (RTA), has proved itself to be highly successful in an experiment carried out at the University of Washington, Seattle [Guan et al., 2006]. A brief interview concluded the exercise.

2.2. The Exercises

The test manager was informed in advance about the intention of the exercises and success criteria were defined for each task. More precisely - we described 10 tasks to be solved. These tasks included some rather general and often used functions on many web sites. Others would necessitate some knowledge of statistical terminology in order to navigate the Statistics Denmark’s web site. In addition, a few more complex tasks were to be solved.

The test was separated into two parts:

– To find information directly on the homepage (www.dst.dk) and use some of the functions available from there (print, search, ..), tasks 1-5.

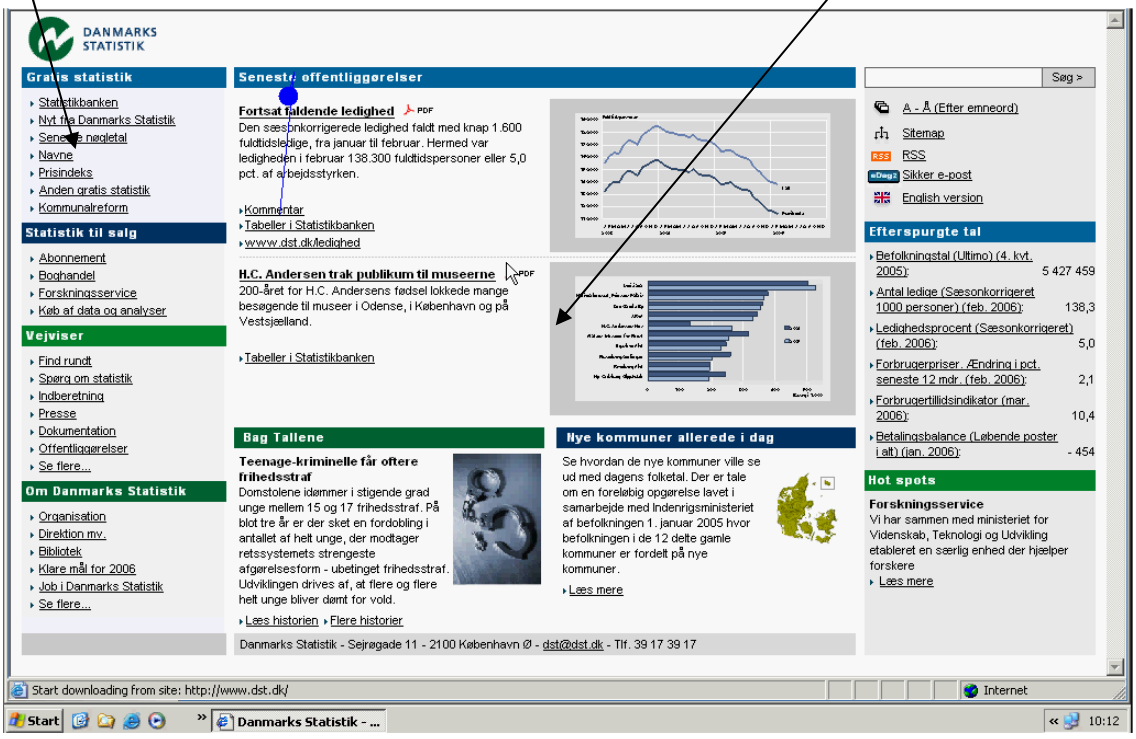
– To extract specific data from the StatBank (www.statbank.dk) and use some of the functions available there (file download, additional documentation, related publications…), tasks 6-10.

3. Test Results

Part one of the exercise involved the more simple tasks. Looking up information on the website was in general possible either through direct links or through the sitemap. Part two - consisting of retrievals from the StatBank - caused more challenges: even though the person being tested considered that he/she had solved the task it was highly probable that he/she had not found the correct information. The follow up interviews would provide the relevant information regarding the problems encountered in this case which included terminology, layout and structure.

In general those being tested managed to solve the tests. However, the time spent to complete an exercise, proved to be highly variable:

Example 1:
Find information about changes in the local authority structure. Print the page.
The information is directly available from the homepage at two places: namely (A) and
(B)



Example 1.

Four students solved the task by selecting the link with the map (A), while three selected the left column (B). The rest chose to go via *Sitemap* in the upper right column. When the time arrived to print the information, eight used the print icon on the site while two used the print icon in the browser.

The picture in example 1 is the first glimpse from a video, showing the first fixation to be in the upper middle part and going downwards (see the blue dot and trace).

The pattern is similar when dealing with the more complex exercises. The video films show that a user can pass the correct link several times without actually noticing it. Several of those being tested failed to observe the relevant links and went instead to *Sitemap* or *Find your way*.

In task two, the population in Denmark was requested. It is considered a simple task and the result is at the front page in the right hand column. 5 students found the information in the correct place, but they considered the information to be a link and not the actual result. The remainder utilised three clicks in the left hand side column (Statistics free of charge → Latest key figures → Population). Comments from the interviews suggest that the information on the left hand side is very compact which causes the focus to be on this side.

In task 3, the usability of our search function was what was required to be tested. It is placed in the upper right corner and 2 students failed to find this search function. In subsequent discussions it was mentioned that they had expected a search function to be placed in the middle of the page. Similar results are recorded in a task where the search involved was to be alphabetical.

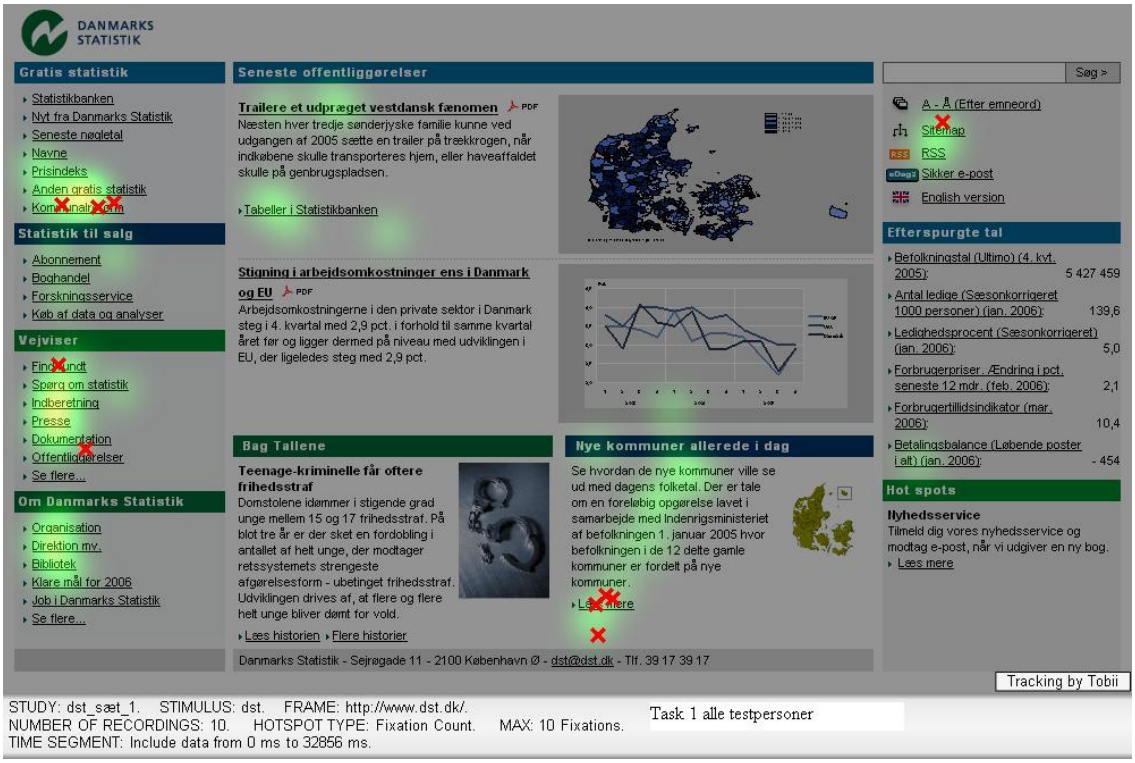
Time in seconds	task 1	task 2	task 3	task 4	task 5	task 6	task 7	task 8	task 9	task 10
Minimum	0:32	0:17	0:29	0:34	0:29	0:57	1:22	0:28	1:07	1:31
Maximum	1:35	0:53	4:03	3:10	4:28	3:47	3:30	2:50	2:59	4:24
Gave up	1				3		1			1

Table 1. Minimum and maximum time needed to solve the tasks

In a test such as this, the users expect there to be a solution and thus will continue to search. However, in real life some users would probably have given up at a much earlier stage as three to four minutes is considered to be a long time on the web.

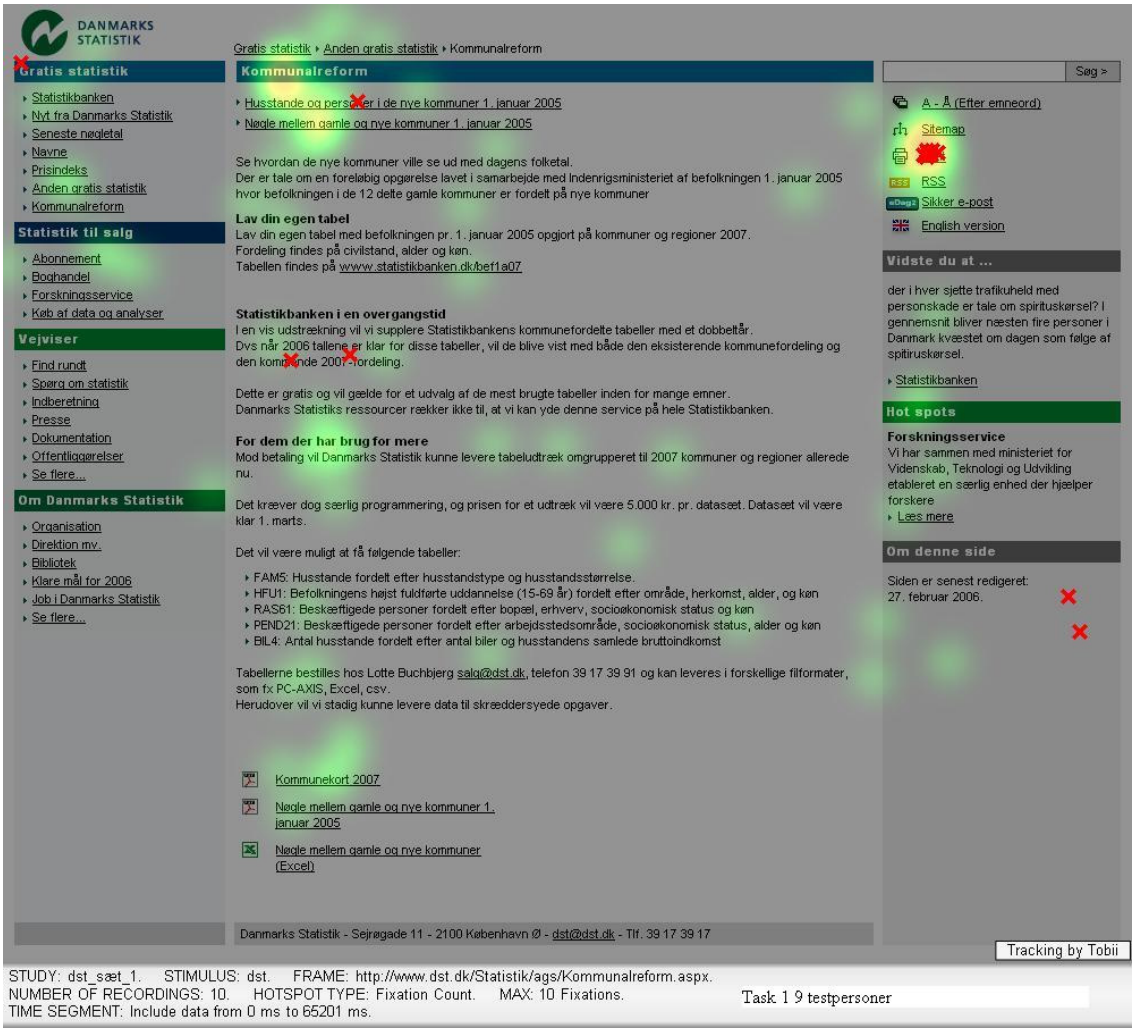
3.1. Heat Maps

Results from the tests can be shown using the so called *heat maps* or *hot spots*. This method will stack all the information concerning a specific exercise from each different user onto one map. However, the data analysis tool, Clear View, is unable to handle “pop-ups”, and is also unable to manage data download whereas the majority of the tasks from StatBank users are on pages that change during their action, for instance when scrolling the value boxes and clicking *view table*. This information becomes meaningless when stacked.



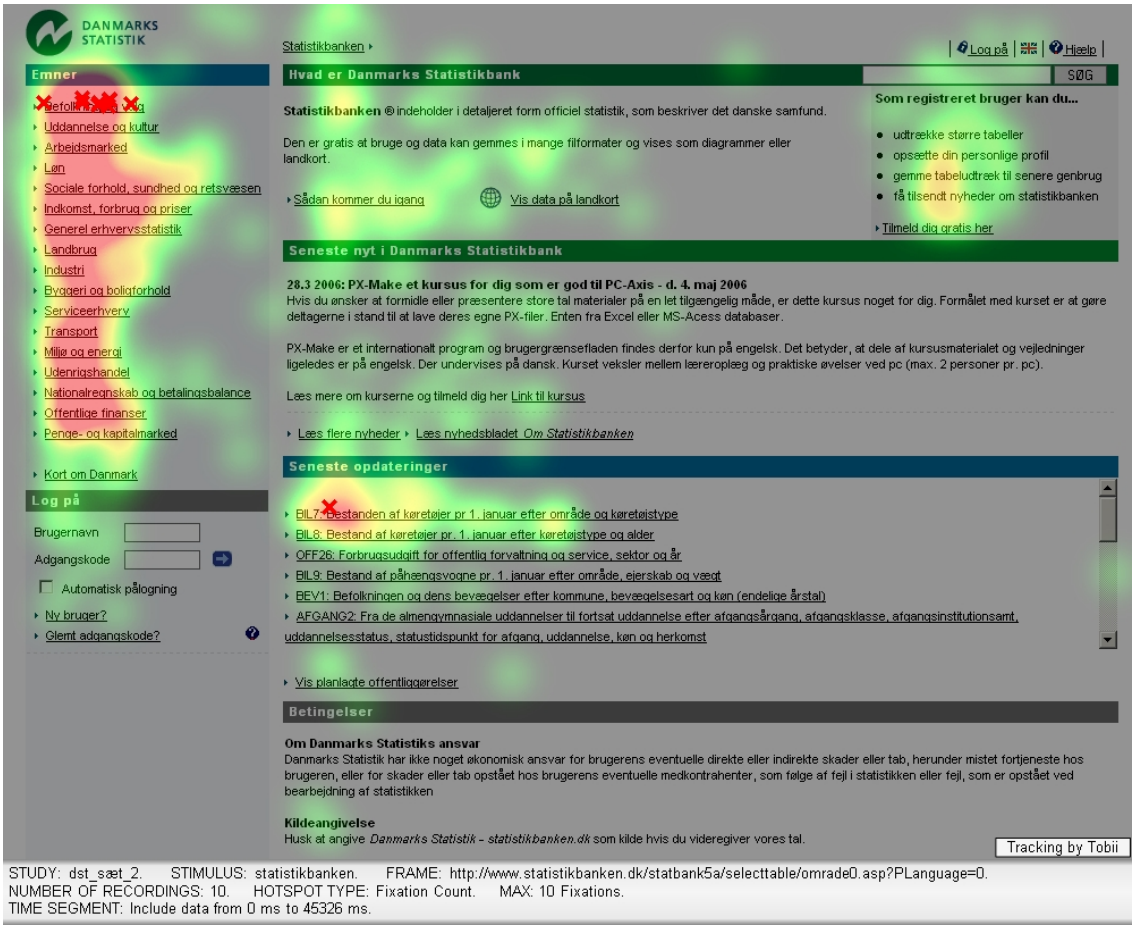
Example 2.

The heat map in Example 2 shows the same exercise as that for the video in example 1. Here all the activity for the ten people tested is combined. A red spot indicates that 10 fixations have been counted in this position (there is no such concentration on this page). The number of fixations decreases through orange and yellow. A light green spot indicates 1-2 fixations. The red crosses mark mouse clicks.



Example 3.

Example 3 shows the second part of task 1. The person being tested should print out the information concerning the municipalities found. The green and yellow spots indicate that the page had been skimmed before the print icon was clicked.



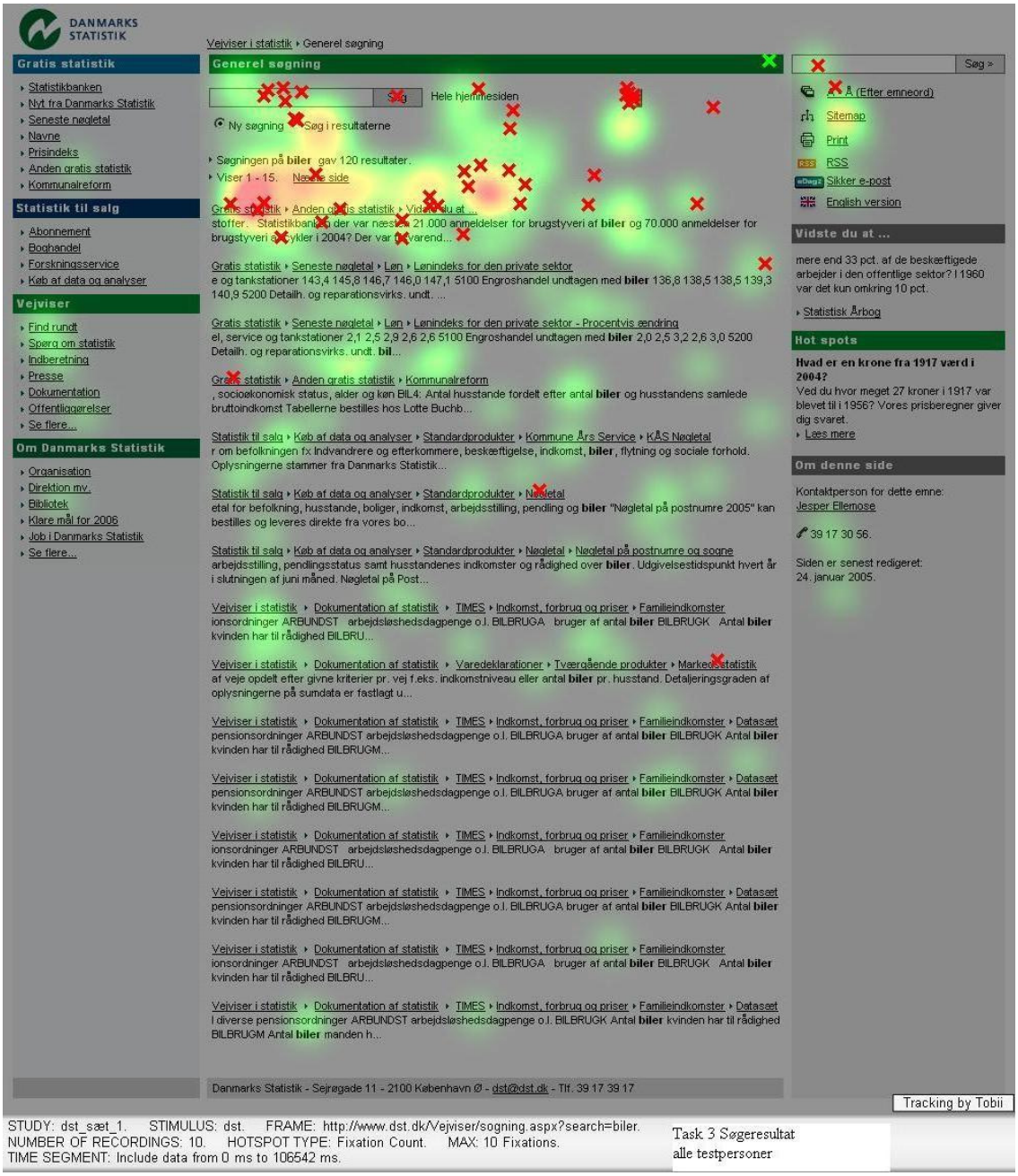
Example 4.

Example 4 is from the StatBank. The task is to find the number of Swedish citizens in Denmark. In this case the navigation is by topic using the left side column. This appeared to be evident to everyone and it appeared that none had considered logging on - which was, in fact, unnecessary.

However, when selecting tables in the following step, the text lines were long and apparently difficult to understand and it proved difficult to determine the difference between them. This was also stated by the testers in the follow up interview.
Many table titles within a specific topic are more or less identical for the majority of the first words and in order to select the correct one, the full title must be examined:

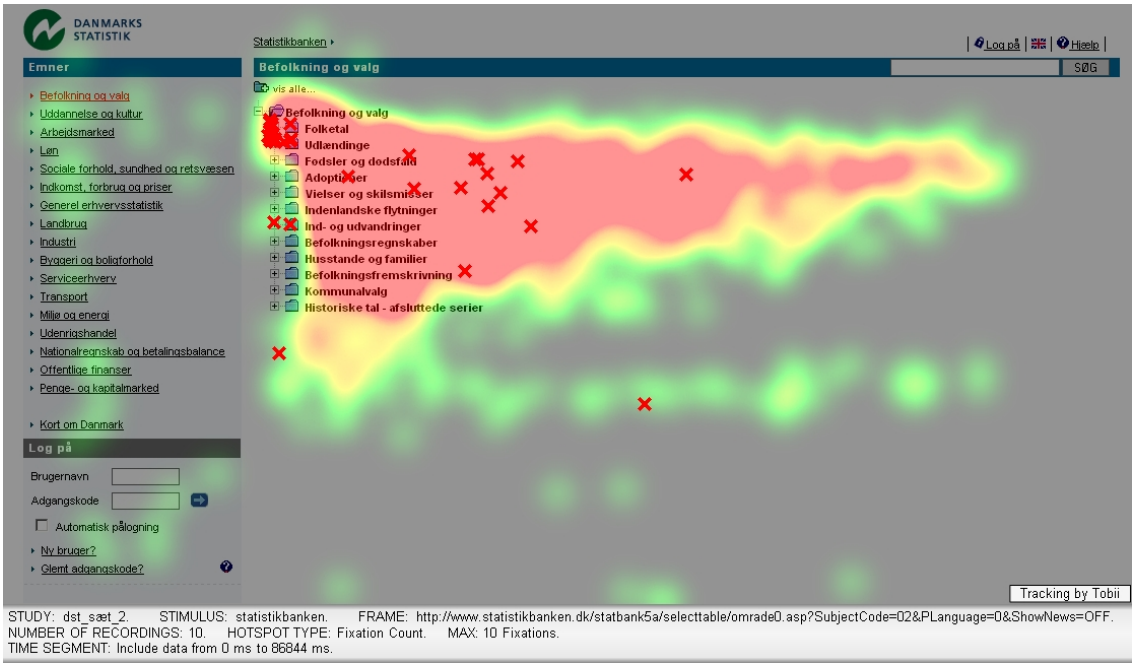
Sickness and maternity benefits, all cases by type of benefit, age, sex, movement, region and time
Sickness and maternity benefits, all cases by type of benefit, age, sex, movement, and time

Example 5 is a result from a search on the web. Only the very first links are looked at and our test confirms – as do some Google tests - that the user does not read much on the page. The first 1-2 hits are read to the end. However, lower down in the list, it is only the first part of the line that is looked at. At the bottom we find only scattered glances. It is important that the most relevant hits are placed at the top of the page as is also the case in Google.



Example 5.

The next requirement was to test the download facility from the StatBank. In previous surveys some users had commented that a download facility to Excel was required and as this functionality already existed, it was decided to test it using eye tracking. The result was quite surprising and encouraging: all those being tested easily found the download function and selected the required format.



Example 6.

In Example 6 the red crosses mark mouse clicks – also on the subsequent pages, where different choices have caused a different number of tables within the folders. This has caused a somewhat chaotic look, which is not very useful for the analysis. It is a weakness of the system which, however, is overcome by the videos.

4. Conclusions

Conclusions will be two dimensional:

- Evaluation of the navigation results
- Evaluation of the method.

4.1. Evaluation of Navigation Results

The test has confirmed several assumptions concerning how users navigate within the site – and has also demonstrated some erroneous assumptions. Firstly, there appears to be some weak points in the structure of the general home page www.dst.dk. The interviews showed difficulties in managing the left side column which is used in navigation. It is very compact with many links grouped within four categories:

– Statistics free of charge
– Statistics for sale
– Guidance to statistics
– Organisation

Within each of these groups 4-7 links are visible from the front page. For instance *StatBank, Researchers Forum, Book shop, Management*,…while others are one more click down and are hidden behind *See more* links. This is a solution that was selected in order to avoid scrolling on the front page. This user friendliness was not in fact appreciated by all users.

Some users felt that what was behind the groups in this left hand column of the page were not evident. None of those being tested chose the *See more* link. Most visitors are looking for statistics; statistics within a specific topic. However, the structure of the site is "product" oriented: *Statistics for free, StatBank, News releases, Statistics for sale* etc. The sheet anchor is the *Sitemap* and *How to get around*: from there users obtain an overview of the site possibilities. The heat maps indicate that nobody looks in the upper left corner where we have placed the logo. Tests from other web sites have shown similar experiences.

4.1.1. Lesson 1 learned

The difference to the StatBank is evident: here the different statistical subject areas are listed in the left hand side. The people being tested thoroughly investigated all the options before making their selection. In a future layout revision of dst.dk, a home page based on subject areas should be considered.

4.1.2. Lesson 2 learned

Another lesson learnt concerns the length of text lines. It is evident when search results are presented and where StatBank table titles are presented in a list that only the first few results are observed and approximately one third of the line. In the case of StatBank table titles, however, the last third of the title is crucial in order to distinguish one table from another.

The search results should be presented in such a way that the content is obvious from the very start of the line. This improvement has already been put into practice. The table titles are (unfortunately) created automatically, however, and we will have to reflect on alternative solutions. One possibility could be to split them into more sub-topics. This could prevent the parts of the table title being repeated, and they could then be found in the title of the sub-topic. Pictures and icons are sometimes seen to identify groups of statistics. However, it is our experience that this only works for themes. It is hard to find pictures that will be interpreted correctly with reference to differentiation at lower levels, for instance between unemployment rates and total unemployment.

4.1.3. Lesson 3 learned

Among the selection alternatives in the left hand side navigation are a few "hidden" behind the text *See more…* Statistics with reference to visits for specific pages had shown that the hidden alternatives were very, very rarely found. The eye tracking test confirmed this. To solve one of the tasks, the person being tested had to unfold the *See more*... Not a single user clicked the *See more…* The Site map solution was chosen instead.

Hidden alternatives in the navigation are relatively useless. What we want the user to see should be visible and comprehensible from the first page and the provision of meaningful labels to the links is the be-all and end-all for the usability.

4.1.4. Lesson 4 learned

In general the left side column appears to be frequented, particularly the upper section. In the mid column the focus appears to be on the upper left section and partly down to the lower right. The right column is not often in focus.

Nothing of importance should be placed in the right side column. Only the upper part of the right side is observed: that is where the small icons are to be found: print, RSS, Site map, language. This is more or less the standard for many web sites.

4.1.5. Lesson 5 learned

Solving the StatBank tasks (tasks 6-10) caused problems to more users. Some of them spent a great deal of time working on the tasks, yet the Help function was not activated by anyone. One could ask: is there a need for more help info or is there a need for more simplicity? Perhaps the majority of the information is entered onto the pages in order for the producer to have a clear conscience, but all it succeeds in doing is to disturb the clarity for the users? A bad structure or layout cannot be repaired by extending the texts.

4.2. Evaluation of the Method

Statistics Denmark has been using different methods to measure user satisfaction:

1. Usability test, inviting the users to "concurrent think aloud" CTA [Wulff, 2002]
2. Pop-up web surveys with the possibility for user comments [Jensen, 2006]
3. Focus groups
4. Analysing search results
5. Hot line support interviews
6. Eye tracking combined with follow-up interviews

In particular, methods 1 and 2 have been used with success for years. However, none of these methods provides the full true picture and all possess advantages as well as weaknesses. The eye tracking can lead our attention to some disadvantages regarding the structure of the site. For instance it completely supports the view that scrolling through long pages should be avoided. A more surprising result is that users accustomed to browser navigation appear to use functionality such as tracking backwards and printing by means of the browser rather than from icons on the specific page. Another example is that it has been the intention of Statistics Denmark to show commonly asked for figures on the front page which is done in the right hand column. The eye tracking nevertheless shows that the users hardly notice the right hand column. In the follow up interview some said that there was too much information on the page to interpret it all.

4.2.1. Weaknesses

It is difficult to know how conscious the user is concerning what the eyes look at: do they watch or do they wonder? That is also why other users of the method recommend it to be combined with a follow-up interview. Confronted with the gaze replays, the user can comment on how the task was tackled.

The version of the analysing tool, Clear View, that was used in this study did not compensate for scrolling, frames and pop-up windows and the aggregation of the results of those tested into a heat map proved to be useless in such cases.

Until such time as the method allows us to separate the different layers, it will not be of significant use in the StatBank, but will still exist on the web site. The development of the software will be followed with considerable interest.

The equipment has, so far, proved to be relatively costly: 12,000-15,000 Euros which is not realistic for one or two annual tests. The results obtained in the Canadian study [Guan et al., 2006] showed that retrospective think aloud RTA gave results close to the eye tracking results - in particular regarding the less complex tasks. RTA could, in such cases, be an economic substitute to eye tracking. However, we believe that the price will drop, as has been the case for most electronics.

4.2.2. Advantages

The eye tracker is discretely integrated into a monitor without any visible or moving "tracking devices". The people being tested are allowed to move freely in front of the screen. This non-intrusiveness ensures that respondents behave naturally. It allows for the collection of eye tracking data simultaneously with standard test procedures, for instance think-aloud usability testing or follow-up interviews. This was also done in our case. The first part of the test (tasks 1-5) was followed up by an RTA, where those being tested draw their recalled navigation on a sheet of paper. There was a rather good agreement between the actual eye movements and their understanding of the behaviour. When there was a disagreement it was usually an opinion that he/she did "better" than was shown by the eye movements. Additionally the person being tested does not feel any incompetence is being revealed.

One of the significant advantages of the eye tracking method is that the results can be readily shared between team members working with the web in addition to providing a convincing document for the managers. The layout of web sites is often an area where there are conflicting individual opinions regarding taste. The videos and the heat maps assist us to focus on how users tend to see and interpret the site.

References

Guan, Z., Lee, S, Cuddihy, E., and Ramsey, J. (2006). "The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking", CHI 2006 Montréal, Canada

Jensen, J. E. (2006). *Brugerundersøgelse af www.dst.dk 2002-2006*, www.dst.dk/brugernesmening

Jensen, J. E. and Wulff, A. (2004). *Brugertest af www.dst.dk/boghandel*, Danmarks Statistik/Statistics Denmark.

Jensen, J. E., Knudsen, L., and Wulff, A. (2002). *Usability test af Statistikbanken,* Danmarks Statistik/Statistics Denmark.

Lund, H. (2006). *Brugertesting via Eye-tracking*, The Royal School of Library and Information Science, Copenhagen 2006

Wulff, A. (2002). *Usability test af Statistikbankens grænsesnit*, Danmarks Statistik/Statistics Denmark.

Wulff, A. (2006). *User satisfaction survey of StatBank 2001-2006*, www.dst.dk/usersurveys, Danmarks Statistik/Statistics Denmark.