Concept Testing Toward a Patient-Validated Information Architecture

Prototype Development of Healthtalk Norway

Tore Høgås
Centre for Quality and Development
University Hospital of North Norway
Tromsø, Norway
e-mail: tore.hogas@telemed.no

Torsten Risør
Faculty of Health Sciences
UiT The Arctic University of Norway
Tromsø, Norway
e-mail: torsten.risor@uit.no

Marianne Trondsen, Line Lundvoll Warth, Kari Dyb and Hege Andreassen
Norwegian Centre for E-health Research
University Hospital of North Norway
Tromsø, Norway
e-mail: marianne.trondsen@telemed.no, line.lundvoll.warth@telemed.no, kari.dyb@telemed.no
and hege.andreassen@telemed.no

***Abstract*—A Norwegian research group is adopting the Database of Individual Patient's Experience of Illness (DIPEx) international methodology standards for collecting qualitative research into people's health experiences and disseminating it on a web site. We are in the concept phase of developing the web site, and decided to build a topical ambiguous taxonomy together with a more clinically influenced taxonomy with top-level labels "Health and lifestyle" and "Illness" for the information architecture of the web site. In this paper, we report from usability testing of the top-level label of the topical taxonomy. We ran qualitative and quantitative A/B tests on wireframe concept sketches. The two top-level labels were a generic variant, "Topic", tested against the control variant, "Everyday life". Both qualitative and quantitative tests indicate better results for "Everyday life" as the top-level label for the topical ambiguous taxonomy of the web site. While not fully conclusive, the results provide reasonable confidence in the more descriptive label "Everyday life" at this early stage. It is preferable in that it both seems to create a more coherent set of expectations amongst the users, and more closely matches the content of the web site. The concept test is therefore deemed a useful first step in a rigorous testing program to ensure that the development process is informed by a patient-validated information architecture.**

Keywords- usability; information architecture; A/B testing; health experiences; qualitative research.

I. INTRODUCTION

The Internet is an increasingly important source of information for health purposes [1]. Evidence suggests that peer-to-peer exchange of health experiences has been one of its most transformational features [2] and is most likely to engage site users [3][4]. However, many web sites and other media sources present only a few anecdotal accounts or are skewed toward heroic or exceptional testimonials. Hence, there is a need for comprehensive research based collection and dissemination of patient accounts that includes the multiplicity of subjective everyday experiences.

To meet this need, the project "Healthtalk Norway" will pilot the Database of Individual Patient's Experience of Illness (DIPEx) methodology of health experiences research [5][6] in Norway, with the aim to "promote excellence of qualitative research into people's experiences of health and illness" [7]. DIPEx, developed by the Health Experiences Research Group (HERG) at the University of Oxford, is a qualitative research methodology based on in-depth interviews with patients and carers, for developing, producing and systematizing knowledge on peoples' health experiences. The core of the DIPEx methodology is a web site disseminating these health experiences to patients, carers, students, health professionals, health care services as well as the general public. Extracts from interviews are presented on the web site as video, audio, or text.

Through the DIPEx International network, researchers in 11 other countries have adopted the methodology, and at the time of writing, there are nine active web sites globally. A research group based at the Norwegian Centre for E-health Research (NSE) and with members from two Norwegian universities (UiT and NTNU) is currently working on a Norwegian Healthtalk web site. The first stage in this work is a feasibility study where one of the activities is to develop a prototype of the web site. Web development is an iterative process of creating and testing, often in four main phases: Concept, Prototype, Build, and Implement. To validate choices in functionality, content, structure, and design, it is important to employ a wide variety of tests throughout all four phases. In the following, we present results from usability testing in the very beginning of this process: concept development.

The starting point for the concept phase was a competitive audit of the nine active DIPEx web sites. The original United Kingdom site healthtalk.org was established by HERG in 2001 and has grown organically over the years, currently comprising more than 80 sections covering various conditions, diagnoses and health topics through approximately 250 interview excerpts. This means that its navigation is comprehensive, but has usability challenges because of scale. The other eight sites, in contrast, have very rudimentary navigational structures where individual diagnoses such as stomach cancer or epilepsy are listed as top-level labels, lacking any overarching categories. Therefore, they may face challenges as they grow. Hence, rather than duplicating existing web sites, we decided to start work on the information architecture for our site anew.

The research question in this paper is in what way quantitative and qualitative concept testing can be a useful first step toward a patient-validated information architecture. Our objective is to present a model for information architecture that will be both suitable and scalable, as well as user friendly. To this end, Section 2 discusses information architecture considerations for this project in the context of expected user categories or idioms. Section 3 describes the method of concept testing in the usability field and how we applied it in our tests. In Section 5, the results from these tests are discussed, and conclusions from these tests are drawn in Section 6.

II. INFORMATION ARCHITECTURE IDIOMS AND CONTEXTS

In the context of web development, information architecture can be described as the discipline of organizing a web site's information so that its users can find the right answers to their questions [8]. The challenge for a web site for the general public is to find organization schemes that are meaningful to a large and heterogeneous target audience, using labels that are relevant to and resonate with the user's own categories. At the same time, these schemes must be robust enough to accurately represent current content and scalable enough that we can reasonably expect them to represent future content.

In our case, there is an additional challenge: Users as patients are known to have several cognitive domains regarding health and present different parts of these domains in different contexts; collectively known as illness idioms [9]. This is often the case even for health care professionals who find that, as patients, their medical knowledge is not as helpful in resolving the relational and everyday challenges associated with being a patient (particularly with a chronic illness). Thus, while at a clinic, specific questions about symptoms and medicine may take center stage in communication, whereas questions about managing cooking, driving a car or using the bathroom may be much more central in a home context.

The content of the web site in question consists of people's health experiences, which encompass both the clinical experiential field and most areas of everyday life. Its categories must therefore reflect the multiplicity of patients' illness idioms. It is our aim to provide an alternative to official health information services and give prominence to patients' everyday experiences. We therefore decided to implement a dual taxonomy organization scheme both in site navigation and in faceted search results to cater to search-dominant users.

First, to facilitate known-item searching we will use *diagnoses/illnesses* and predefined *health and lifestyle* topics as the primary taxonomy. This approximates an *exact organization scheme*, as there is a large degree of consensus around diagnostic classification. Moreover, this taxonomy is already in use at the official Norwegian health portal helsenorge.no. The Norwegian Directorate of Health has based the diagnoses/illnesses part of the taxonomy on the Medical Subject Headings (MeSH) controlled vocabulary and used a variety of methods, including usability testing and web statistics analysis, to increase its usability for the Norwegian public and adding the health and lifestyle dimension. Additionally, since official health information is increasingly routed through this portal, many in the target audience will be familiar with this taxonomy when exposed to it on our web site.

In addition, we decided on supplementing the primary taxonomy with a *topical ambiguous organization scheme*. This has the disadvantage of adding cognitive load for the user, i.e., added mental resources required to use the site [10]. However, it was judged necessary to account for the fact that so much of the information we want to convey from patient experiences does not fall within the clinical field as defined in the primary taxonomy, but is part of a different idiom for the same illness. Examples of such topics range from how one deals with getting a diagnosis to how one's illness or health condition affects sex and intimacy. Additionally, this will support a common serendipitous mode of searching where the user has not necessarily formed a clear idea of what she is looking for.

When designing a topical organizational scheme it is crucial to develop a typology that has a strong topical relevance relationship to its content. The stronger the topical relevance, the lower the cognitive load for the user. Topical relevance can be arrayed in three facets: the functional role of information, how information contributes to the user's reasoning about a topic, and how information connects to a topic semantically [11]. When evaluating the topical relevance of a term in a taxonomy, we have to evaluate all three facets.

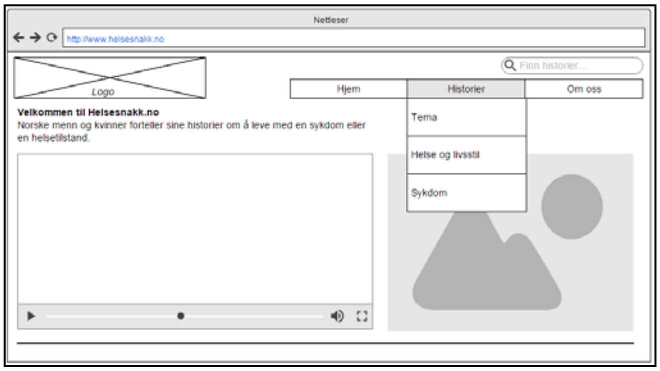
III. TOP-LEVEL TOPIC LABEL TESTING

Topical relevance is more important in the top level of a taxonomy than further down in the hierarchical structure, since the top-level label often also serves as a user's navigational entry point. For the exact organization scheme, we appropriated the top-level labels from helsenorge.no as "Sykdom", or "Illness" in English for the MeSH-based structure, and "Helse og livsstil" ("Health and lifestyle") for other categories such as pregnancy, nutrition or smoking.

We then needed to determine which top-level label would be the most appropriate for our topical ambiguous organization scheme. The first candidate was the generic label "Tema" ("Topic" in English). This label was judged strong in the facet of functional role, and has the advantage of flexibility through being generic. The other was "Hverdag" ("Everyday life" in English). While less flexible, it was judged stronger in both the user reasoning and semantic facets of topical relevance. However, it was not a clear-cut decision as to which candidate would ultimately have the highest topical relevance.

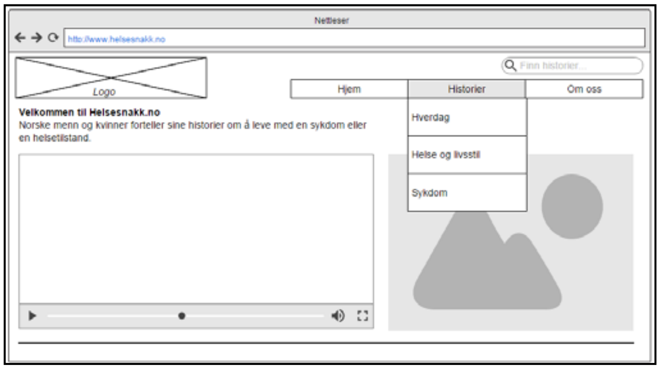
To find out in practice which candidate would result in the least amount of added cognitive load for users of the site we decided to run qualitative and quantitative A/B tests on wireframe concept sketches. An A/B test is a randomized experiment where users are exposed to one of two variants of a web site design. The variants are identical except for the one variation that is being tested, in this case the top-level label of the secondary taxonomy. Figure 1 is the wireframe of the generic ("Topic") variant we tested, formally the control variant, i.e., the null hypothesis for statistical evaluation. Figure 2 is the wireframe of the treatment A ("Everyday life") variant.

As this is only the first phase in a longer development process of iterative designs, we used the Notable web-based test platform to conduct remote testing. For each A/B test, we only need to recruit a single pool of test users, and the platform itself randomizes which of the variants is shown to the users.



Velkommen til Helsesnakk.no
Norske menn og kvinner forteller sine historier om å leve med en sykdom eller en helsetilstand.

Figure 1. Wireframe of the variant with the "Tema" ("Topic") label.



Velkommen til Helsesnakk.no
Norske menn og kvinner forteller sine historier om å leve med en sykdom eller en helsetilstand.

Figure 2. Wireframe of the variant with the "Hverdag" ("Everyday life") label.

A. Qualitative A/B Testing

For the qualitative test we recruited users via email, phone calls, and social media, limiting self-selection bias by recruiting users of both genders, of varying ages and levels of education, and excluding users working in either the health sector or technology/new media. As the web site in question has a large target audience, i.e., the general public of Norway, additional profiling criteria were not deemed necessary for either of the two tests. The recruitment method, which was largely online, as well as the online test delivery itself, ensured a certain minimum of Internet expertise in the user pool.

The minimum recruitment pool was set at five users per variant, following Jakob Nielsen's findings, which indicate that usability testing yield the best results when conducted in an iterative process with only five users in each test [12].

The test started with contextualizing the task: "You will see a sketch of a new web site with patient experiences. The site navigation will have three categories of experiences, and we would like your feedback on one of these options." The test then showed users the wireframe sketch with the label we were testing for highlighted, and the instructions: "Consider the highlighted menu option on this sketch. What do you think you would see if you clicked on that option?" We designed this open-ended question to determine topical relevance from the extent to which their responses match the intended meaning of the label.

B. Quantitative A/B Testing

For the quantitative test, we used social media for self-recruiting, i.e., asking a large number of people to volunteer for testing without applying any kind of selection criteria. Self-recruiting tests inherently run the risk of self-selection bias toward users who are more Internet-savvy, especially when online. However, since A/B tests are a form of multivariate research, this risk does not apply [13]. Even if responders are above-average experienced Internet users, their bias applies equally to the two variants, so we still get meaningful data.

The minimum requirement pool was set at twenty users, again following recommendations from Jakob Nielsen based on the findings that testing with twenty users gives you a confidence interval of maximum +/- 19%, while you would need as many as 76 users to reach +/- 10%. [14]

While the qualitative test was designed to determine topical relevance through users' written expression of reasoning about the label, the quantitative test focused on behavior in interacting with the label. Thus, while the contextualizing introduction was similar, the task and instruction were different. Test users were shown the wireframe sketch without any markings and the instructions: "Imagine that you have a serious illness and need a practical question answered. Look at this image. Click where you would have clicked to find information on how this illness affects driving a car."

IV. TEST RESULTS

A. *Results from Qualitative A/B Test*

The qualitative test received six responses for each variant, one more than the minimum requirement. The responses are translated and reproduced below with original punctuation preserved.

Responses to the control variant, i.e., what respondents expected to find under the "Topic" label:

1. What kind of illness
2. Heart-warming stories
3. Various diagnoses? (not obvious)
4. Gender
5. Meeting the doctor, monitoring after illness
6. Symptomless. Now what?

Responses to the treatment variant, i.e., what respondents expected to find under the "Everyday life" label:

1. How the illness has changed my everyday life?
2. How the illness affects my daily life
3. This is how I live with my illness
4. Living with the illness in my everyday life, 'trying' to live a normal life
5. This is how I feel
6. (Blank response)

B. *Results from Quantitative A/B Test*

The quantitative test received a total of 62 responses; 32 for the control variant and 30 for the treatment variant. These results are summarized in the contingency tables below. We have defined "success" as a click on the label we were testing for, "Topic" and "Everyday life" respectively. Our definition of "failure" is any other click on the image. Table 1 shows all collected results. Table 2 displays results filtered on responses where the user spent more than 10 seconds looking at the image before clicking.

TABLE I. ALL RESULTS

	"Topic"	**"Everyday life"**	**Marginal Row Totals**
Success	10	14	24
Failure	22	16	38
Marginal Column Totals	32	30	62

TABLE II. SPEED FILTERED RESULTS

	"Topic"	**"Everyday life"**	**Marginal Row Totals**
Success	3	9	12
Failure	19	20	39
Marginal Column Totals	22	29	51

V. DISCUSSION

A. *Qualitative Analysis*

Out of six responses to the "Topic" variant, three were irrelevant to the planned content of this taxonomy: "What kind of illness", "Various diagnoses? (not obvious)", and "Heart-warming stories." The two first are obvious misunderstandings of what might be under the "Topic" label, since there already is a separate top-level label for "Illness." One of them calls attention to the user's uncertainty by adding the question mark and the parenthesis stating explicitly that the label is not obvious. The third answer seems to expect the type of content other sites skew towards; heart-warming stories of heroic endurance. The sites built with the DIPEx methdlology are not about such stories; they are about the unheroic and unfiltered experiences of regular people.

The other three responses to this variant were relevant to varying degrees. Responses 5 and 6 in particular go beyond the clinical experiential field to what happens after an illness. Nevertheless, in sum these six responses indicate that the flexibility of the "Topic" label turns to plasticity, with widely varying understandings of its meaning.

Turning to the six responses to the "Everyday life" label, we see that with the exception of the one blank response, they are all relevant to the intended meaning of the label. Three of them are even paraphrasing the label: "How the illness has changed my everyday life?", "How the illness affects my daily life" and "Living with the illness in my everyday life, 'trying' to live a normal life." As many as four responses make an explicit connection between this label and the adjacent "Illness" label of the other taxonomy, which indicates that the users' reasoning is that the content behind this label is connected to, but different from, illnesses and diagnoses. Indeed, the interplay between the three top-level labels seems to be much more productive when the "Everyday life" label is used, than with the "Topic" label.

It is also valuable input that two of the respondents to this variant introduces the subject "I" in their responses: "This is how I live with my illness" and "This is how I live." In combination with the leading demonstrative pronoun "this", it indicates that they have formed a quite strong identification with the level on a personal, emotional level. The same may be said of the response that contrasts "living with the illness in my everyday life" and "trying to live a normal life." A diagnosis often represents a biographical disruption in the patient's life, and we know that restoring normality and a coherent biography in times of illness is demanding emotional work [15].

The fact that four out of twelve responses in total are irrelevant or blank may indicate that they did not fully understand the task, which is a weakness in unmoderated online tests. Nevertheless, this weakness is the same for both variants, yet there is a clear difference in responses to two variants tested. While users' free-form responses to the "Topic" label go in different directions, the "Everyday life" label elicits responses about relationships between illness and daily life, with indications of both an emotional component and changes induced in the patient's life. The label “Everyday life” thus creates a more coherent set of expectations in users along the facets of both users’ reasoning about the label and its semantic value. These expectations more closely match the content that we will publish on the web site, resulting in higher topical relevance.

Seen as a whole, the results from the qualitative test also provides valuable insight into what kind of content prospective users would expect or even want from our web site. These expectations cover the entire spectrum from first meeting the doctor and getting a diagnosis, through the illness progression, and beyond.

B. Quantitative Analysis

The results from the quantitative test corroborate the indication that the “Everyday life” label has higher topical relevance for users tested. From the results in table 1 we can calculate a success rate of *0.31* for the control variant, while the treatment variant has a success rate of *0.47*. The difference in success rates is *0.16*, which is clearly in favor of the treatment variant; “Everyday life.”

The speed filtered results in Table 2 are of interest based on the assumption that variance in cognitive load affects slow users, defined as users spending more than 10 seconds on deciding where to click, more strongly than fast users. In other words, if we are to reduce cognitive load as much as possible, responses from slow users are more important than responses from fast users, because its effect is amplified.

These results show that the control variant has a success rate for slow users of *0.14*, while the treatment variant has a success rate of *0.31*. As expected, success rates are lower for slow users of both variants, yet the difference in success rates is slightly greater for slow users at *0.17*. Thus, both contingency tables indicate higher topical relevance for the “Everyday life” label.

However, there are two caveats to this indication. First, if we use the Fisher exact test to calculate the p-values for the contingency tables from the quantitative A/B test, they are *0.297548* for the full results and *0.192494* for the speed filtered results. In both cases, the values are higher than the significance level of 0.1. or 10%. Therefore, while at least the speed filtered results are within the +/-19% confidence interval we have deemed acceptable for this test, they are technically not statistically significant.

Second, an analysis of the failure clicks on click maps generated by the Notable test platform shows that there are more responses to the control variant that cannot be categorized as navigation-dominant behavior (i.e., focusing on navigational elements) or search-dominant behavior (i.e., focusing on the search box). Figure 3 shows the click map for the control variant with the “Topic” label. In contrast, figure 4 shows a more focused click map for the treatment variant with the “Everyday life” label. While the control variant received 11 clicks that are not on a navigational element or the search box, the treatment variant received only 4.

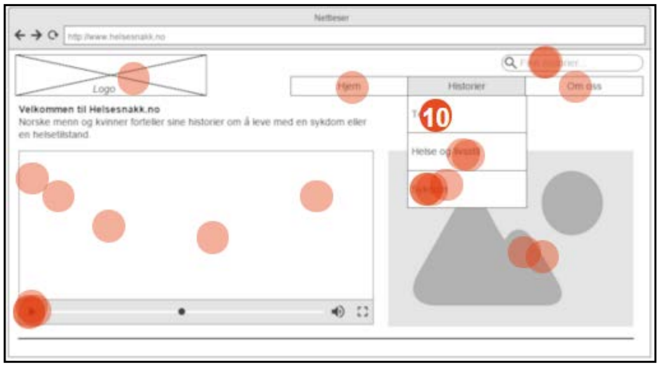

Nettleser
http://www.helsesnakk.no
Logo
Hjem
Historier
Om oss
Velkommen til Helsesnakk.no
Norske menn og kvinner forteller sine historier om å leve med en sykdom eller en helsetilstand.
10
Helse og livsstil

Figure 3. Click map of the “Topic” variant.

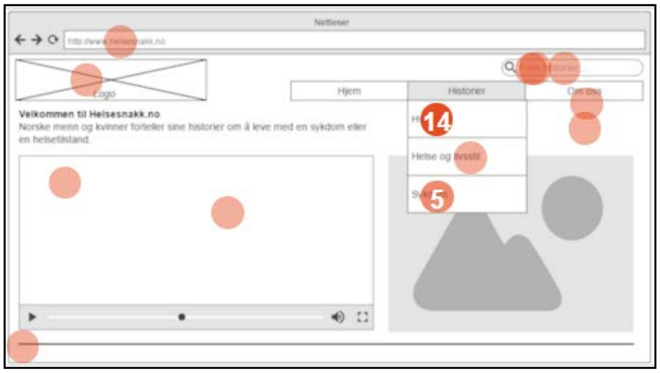

Nettleser
http://www.helsesnakk.no
Logo
Hjem
Historier
Om oss
Velkommen til Helsesnakk.no
Norske menn og kvinner forteller sine historier om å leve med en sykdom eller en helsetilstand.
14
Helse og livsstil
5

Figure 4. Click map of the “Everyday life” variant.

TABLE III. RESULTS WITH OUTLIERS EXCLUDED

	“Topic”	**“Everyday life”**	**Marginal Row Totals**
Success	10	14	24
Failure	11	12	23
Marginal Column Totals	21	26	47

A reasonable hypothesis for behavior that is neither navigation-dominant nor search-dominant in a task such as this is that these users did not fully understand the task. Their behavior does not correspond with the expected patterns of users who are actually trying to complete the task as given. If this is the case, we may have to define all these responses as outliers and exclude them from the contingency tables. Table 3 shows this modified table of results, with failures defined as outliers excluded.

From these results, we can calculate a success rate for the control variant of *0.48*, while the treatment variant has a success rate of *0.53*. Although the latter still has a higher success rate, the difference is minimal at *0.06*.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have described an early web-based A/B test set of information architecture concept sketches. We found an indication in both quantitative and qualitative tests that “Hverdag” (“Everyday life”) is most likely a better choice than “Tema” (“Topic”) as the top-level label when creating a topical ambiguous taxonomy for a patient-oriented web site. The data suggest that this label has a higher topical relevance and therefore results in lower cognitive load for users.

We note that this indication is tempered by the two caveats discussed above. The collected data from the quantitative test is not statistically significant, and excluding outliers reduces the difference between the two variants dramatically. Regardless, the qualitative test gives a much stronger indication since it gives us insight into users' reasoning about the top-level labels. In fact, that is why qualitative methods are often preferred in usability testing; because they do not only show that a given design is problematic, but provide insights into *why* there is a problem as well as *how* you can solve it [16]. Therefore, although test results are not fully conclusive, they have some value in providing reasonable confidence in the "Everyday life" top-level label at this early concept stage.

Going forward in the web development process, we will continue to employ a variety of quantitative and qualitative test methods to achieve patient validation of our information architecture and design decisions for the "Healthtalk Norway" web site prototype. While this first set of tests was delivered online for speed and convenience, we recognize the limitations of this delivery method and will in the rest of the process conduct usability tests in person as well. This experiment has shown that it is necessary to ensure that respondents understand tasks and to improve response registration.

This rigorous testing program will be of importance for further refinement of our model for online presentation of qualitative research on people's health experiences.

REFERENCES

[1] H. K. Andreassen et al, "European citizens' use of E-health services: A study of seven countries," BMC Public Health 7, 2007, p. 53.

[2] S. Ziebland and S. Wyke, "Health and illness in a connected world: How might sharing experiences on the internet affect peoples' health?" Milbank Quarterly, 90(2), 2012, pp. 219-249.

[3] E. F. France, S. Wyke, S. Ziebland, V. A. Entwistle, and K. Hunt, "How personal experiences feature in women's accounts of use of information for decisions about antenatal diagnostic testing for foetal abnormality," Social Science & Medicine, 72(5), 2011, pp. 755-762.

[4] E. Sillence, P. Briggs, P. R. Harris, and L. Fischwick, "How do patients evaluate and make use of online health information?" Social Science & Medicine, 64(9), 2007, pp. 1853-1862.

[5] C. Pope, S. Ziebland, and N. Mays, "Analysing qualitative data," BMJ 320, 2000, pp. 114-116.

[6] S. Ziebland and A. McPherson, "Making sense of qualitative data analysis: an introduction with illustrations from DIPEx (personal experiences of health and illness)," Medical Education 40(5), 2006, pp. 405–414.

[7] H. K. Andreassen, K. Dyb, M. V. Trondsen, and L. L. Warth, "Study protocol: Health talk Norway" Proceedings of the 13th Scandinavian Conference on Health Informatics, June 2015, Tromsø, Norway.

[8] L. Rosenfeld and P. Morville, Information Architecture for the World Wide Web. O'Reilly & Associates, pp. 22-31, 1998.

[9] M. B. Risør, "Illness explanations among patients with medically unexplained symptoms: different idioms for different contexts," Health, vol. 13 no. 5, Sept. 2009, pp. 505-521, doi:10.1177/1363459308336794.

[10] K. Whitenton, "Minimize Cognitive Load to Maximize Usability." Available from: http://nngroup.com. 2013.12.22. [retrieved: 3, 2016]

[11] X. Huang, "Developing a Cross-Disciplinary Typology of Topical Relationships as the Basis for a Topic-Oriented Information Architecture," 20th Annual ASIS SIG/CR Workshop: Bridging Worlds, Connecting People: Classification Transcending Boundaries, 2009, doi: 10.7152/acro.v20i1.12884.

[12] J. Nielsen, "Why You Only Need to Test with 5 Users" Available from: http://nngroup.com. 2000.03.19. [retrieved: 3, 2016]

[13] R. Dewey, Introduction to Psychology. Wadsworth Publishing, 2004. Available from: http://www.intropsych.com. [retrieved: 3, 2016]

[14] J. Nielsen, "Quantitative Studies: How Many Users to Test?" Available from: http://nngroup.com. 2006.06.26. [retrieved: 3, 2016]

[15] M. Bury, "Chronic illness as biographical disruption," Sociology of Health and Illness, vol. 4, no. 2, 1982, pp. 167-182.

[16] C. Rohrer, "When to Use Which User-Experience Research Methods" Available from http://nngroup.com. 2014.10.12. [retrieved: 3, 2016]