Evaluating inventions in locative media

Anders Fagerjord

University of Oslo

Paper prepared for ISMI 2014

Available from http://fagerjord.no/blog/ismi2014

Abstract

Media design can be used for research purposes if it includes a clearly defined research question, and clear evaluation to see whether an answer to the research question has been found. Using a project with locative media for classical music communication as our example, we discuss common evaluation methods from the User Experience field, observing that they all tend to test "interface" and not "content." Instead we propose three other methods of evaluation, that have a basis in humanist theories, such as textual analysis and genre studies: (1) Qualitative interviews with evaluators after the evaluation, asking them to describe the service in their own words, followed by a semantic analysis to get at how they have understood the service. (2) Comparative "experiments," testing alternative versions that are different in key aspects. (3) Peer review by experienced design researchers, who are likely to have a more fine-tuned vocabulary to express their opinions in.

Evaluating inventions in locative media

Every day, new texts are written, new web sites designed, new apps for mobile devices launched and new programs coded. All of these are new, but only some are perceived to be research or innovations. Research for innovation is common in the sciences: Researchers work towards better medicines, machines, instruments, materials, chemical procedures and building techniques. Research for innovation is increasing also within media studies (Bolter, 2003; Fagerjord, 2012; Liestøl, 1999; Liestøl, 2006; Moulthrop, 2005; Nyre, 2014). We should, however, expect research for media innovation to be different from regular design or journalistic practice.

According to Hevner, March, Park, and Ram (2004), design science builds new artefacts from a "knowledge base" of foundations and methodologies, and the resulting design adds to this knowledge base (80). When we create new experiences, services, and genres, we draw on humanist knowledge of genres, storytelling, rhetoric, visual culture, and much more. Can we give back to these disciplines, using design as a research method in the humanities? Following Hevner et.al., I have earlier argued that good design research for media innovation needs a clearly stated research question and rigorous evaluation of the finished product to see what answers are found the research question (Fagerjord, 2012).

This motivation for this article is the design project «Church music in Rome», where we have developed a web site for communicating classical music via mobile devices (Fagerjord, 2011). Using JavaScript to access the phone's geolocation sensors (usually a combination of GPS, GSM triangulation and WiFi triangulation) the device's location is determined. Then it gives a list of six (in the current version) of Rome's historical churches, and the distance to each

of them. The churches are also drawn on a map (see figure 1). For each church, there are two or three audio tracks meant to be played inside (figure 2). Music written for that church is played back, and a narrator explains a little of the music and compares it to the church's architecture and the art inside. Each commentary lasts about a minute. (for more discussion of this service, see (Fagerjord, 2011)).

Blending genres from radio and tourist guides we have created a genre prototype, and hopefully learned more about what a good location-based texts is. Is this design research, or just design? As we move into the evaluation phase of this project, we need to find the right evaluation methods to answer our research questions. Several methods are well established to test the usability of physical and digital products. User experience, understood as a combination of usability, utility, and hedonistic quality also has a range of methods that are agreed to be useful (Hartson & Pyla, 2012). In our work in genre design, however, we have found that these methods are to coarse to give insight into the users' experience of the text in a locative service.

In earlier tests of our application in Rome, we used observation, “think-aloud” methods, semi-structured interviews and a simple survey, methods adapted from the Human-Computer Interaction (HCI) field. We were able to make the interface easier to use, but observational methods hardly gave any insight into how participants experienced the churches together with the music and the commentaries. To find this out, we interviewed and surveyed the users, who responded they liked our application. Survey scores were all positive. We were encouraged, but what could we add to the knowledge base? In the explanatory audio in the application we have, for example, taken care to point out *synaesthetic parallels*, drawing attention to structural similarities in a church’s architecture and music written in the same period. Can we now conclude that synaesthetic parallels is a general principle that works for situated sound? No. We

can't even be certain that it worked in this case: The users may very well have liked other aspects of the application.

In this article, we will discuss the strengths and weaknesses of different evaluation methods that are used in design of experiences, media, and genres, and then propose some new approaches, more firmly based in the humanities, that we will use in the summative evaluation of our proposed new genre.

Testing genre design

In Klaus Krippendorff's (2006) words, "designers create and work out realistic paths from the present towards desirable futures and propose them to those who can bring a design to fruitition" (p. 29). While science is the study of what is, either in nature or society, design is a proposal of what can be made. "In other words, scientists are concerned with explaining an observable world, designers with creating desirable worlds, and statements about either of these worlds call for vastly different methods of validation" (p. 261).

On the other hand, disciplines like engineering, computer science, information systems, pharmaceutics, or medicine also create artefacts belonging to desirable futures while being closely tied to science. These ties are of two kinds: First, the construction of artefacts relies on theories created by (observational) science. Second, the effectiveness of the artefact (the validity of the desirability claim) is tested using similar methods to those used to create the theories, mainly observations and statistics (March & Smith, 1995). Bruno Latour and Steve Woolgar (1986), e.g., recounts how advanced machinery found in a biology lab in 1976 was created within different sciences, relying on earlier results and theories in the same sciences (although in different fields).

If we are to emulate this in genre design, there seems to be two principles that are to be followed: (1) When prototypes of a design are created, we can evaluate them with observational methods, in the same way as it is done in engineering, computer science, or medicine. (2) When designs are based in theories, we can use the same methods that were used in creating the theories to validate the designs. In genre design, these are likely to be genre theories, which are made by close textual analysis of a large number of texts. A textual analysis of the new genre could be a way of validating the design. We will discuss the two principles in turn, beginning with observational methods from the sciences, mostly from psychology.

Observation

The most basic form of evaluation of new designs is observation of use. Evaluators are asked to try out certain features of the new artefact, while members of the design team observe them. Computer applications are often tested in a usability lab, equipped with a one-way mirror more observers can hide behind, and video cameras recording both the user's movements and what happens on the screen (Hartson & Pyla, 2012).

Locative genres such as Church music in Rome can hardly be evaluated in a laboratory. It is in their very nature that they are made to be experienced in a certain place, so evaluators must be taken to the place in question. The main benefit of the proposed genre was also not the interface, but the style of presenting information. The interfaces had to be usable to be sure, and user observation, especially of critical incidents (Andersson & Nilsson, 1964; Flanagan, 1954; Hartson & Pyla, 2012) contributed much to this. But when users were able to access the information in the applications, it was very little to be learned from observing their reaction to what was presented to them. Those who tried out the Church music in Rome app walked around

listening, with no expression of whether they liked what they heard, or if they found it boring, difficult, or interesting, but too long. Other methods are needed to know what goes on in the heads of readers and listeners.

One solution to this is the so-called “think-aloud” test, or *protocol analysis,* where Evaluators are given tasks to solve, and instructed to "think aloud" while performing the tasks, telling the observers how they think and what strategies they use to solve the questions (Lewis, 1982). It has become the most common way of testing computer interfaces, and was popularized by Jakob Nielsen and Steve Krug among others (Nielsen, 2000; Krug, 2010). According to Hartson and Pyla (&Hartson and Pyla, 2012, #55632), "the think-aloud technique is also effective in assessing emotional impact because emotional impact is felt internally and the internal thoughts and feelings of the user are exactly what the think-aloud techique accesses for you" (440). Krippendorff (2006) on the other hand, points out that a known limitation of this method is that many tasks are made automatically in real life, and that verbalizing them slows them down, or may even impair on the respondent's ability to perform them (p. 226).

In a study that can serve as an example of this method, Nielsen and Loranger asked 69 participants a set of about 15 tasks for their usability study of a wide range of web sites (Nielsen & Loranger, 2006). Of these only six questions can be said to concern the "content" of web sites; the information contained in text and images, the style of the prose, and so on. All of them ask for what Nielsen and Loranger call "informational value," in questions such as "list the two main causes of…" or "find out why…". The questions resemble school homework, in fact. Nielsen and Loranger do not ask respondents to evaluate aesthetic qualities or the experience of reading. Still, many of the verbatim quotes from evaluators reproduced in Nielsen and Loranger’s book show

important insights into how readers react to texts, although they are mainly complaints about pages users do not understand or find tedious to read.

It should also be noted that Nielsen and Loranger's preferred method was comparative: They compared web pages to other web pages. Several tasks were web wide, asking evaluators to surf the net for answers. This method could not be used for the “Church music in Rome” project. Like most locative web sites and other genre experiments, it is unique, so similar alternatives do not exist. The researcher may create alternative solutions, however, asking evaluators to think aloud while using versions that differ in important aspects, and then analyse the differences in their comments. This will be expanded below.

In the “Church music in Rome” project, we did use the think-aloud method when testing the navigation system. Users were asked to use the application to locate the nearest church in the program, and to find their way there, thinking aloud when reasoning. This was a helpful technique, and we discovered several improvements to the interface from this evaluation.

Thinking aloud isn't always practical, or even possible, however. We tested our app inside churches, where continued discussion could disturb devoted church visitors. Evaluators were also listening to music and spoken commentary, and talking aloud would make it impossible to listen carefully, thus spoiling the experience they were about to test.

Asking the users

Another, more indirect observation method is the survey. Distributing a survey to evaluators after they have tried out a new artefact is not an observation of their use as such, but it is a way of making their experiences observable, and, perhaps more important, quantifiable.

Experiences are translated into a few categories, and frequencies in each of these categories are summed up and analyzed statistically. Surveys are a way of measuring using a common yardstick, allowing for comparison between tasks.

In design, survey evaluation is contested. Hartson & Pyla have contended that a "questionnaire is the primary instrument for collecting subjective data from participants in all types of evaluation" (Hartson & Pyla, 2012 p. 444). Krippendorff, however, stated bluntly that "structured interviews and questionnaires are the least informative methods of validation" (Krippendorff, 2006).

As mentioned earlier, we created and distributed a simple survey to our evaluators in the first round of evaluation of "Church Music in Rome." We asked the evaluators to rate the service on a survey where they judged 11 questions on a 5-point Likert scale (see Appendix). To distribute it to only five evaluators hardly yields any statistical power to our research, but our intention was more of a pilot study; to see if this survey would give important insights.

What was most striking was that they all gave a 5 (strongly agree) to the statement "It was exciting to be present where the music was first played". We in the design team felt this as a strong encouragement to continue the project. There were less unison feedback to questions about the combination of music, music history, art, and architecture, but as all the users were positive towards the service, we have interpreted this to mean that the synaesthetic parallels work as intended. Still, we could not get rid of a feeling that we might just be looking for indications that the users liked what we hoped they would like. That they have similar tastes as we do, and that the system we are proud of can be considered a success. But popularity is not success in research, knowledge is. An average score does little to advance our knowledge of new genres in location-based media. From this experience, it becomes clear that we must stand with

Krippendorff in his view on surveys; they give little insight into the actual experience of a new design.

More sophisticated surveys than ours exist. Psychologists have in recent years investigated what they call *emotional impact* or *hedonic quality,* such as how appealing the user finds a product's look and feel. AttrakDiff is one questionnaire created to measure hedonic quality (Hassenzahl, 2000; Hassenzahl, 2001). Its authors have tested it statistically and found it valid, but remind us that while the questionnaire measures how pleasurable a product is, it cannot say *what* about the product that creates pleasure or indifference. It is based on a model of user experience where "appeal" is seen as the combination of "ergonomic quality" and two forms of emotional impact, called "stimulation" and "identity". The three kinds of quality, as well as the combined "appeal" are measured using semantic differentials (Osgood, Suci, & Tannenbaum, 1957). Respondents are asked to place their opinion of the product on a seven-point Likert scale between two adjectives, for example,

Pleasant _ _ _ _ _ _ _Unpleasant (measuring Appeal)

Stylish _ _ _ _ _ _ _ Tacky (measuring Identity)

Dull _ _ _ _ _ _ _ __Captivating (measuring stimulation)

Seven semantic differentials are given for each dimension, making 28 differentials presented in random order and polarity. After the test, scores are summed up, and the average is calculated for each dimension. Several statistical tests have been performed on datasets from this questionnaire, and the researchers have found that the scales measuring ergonomic quality and hedonic quality are distinct, and that both contribute to the appeal (Hassenzahl, 2001).

A questionnaire like AttrakDiff is easy to administer to evaluators, is quickly done, and the designers may get feedback on whether the artefact is usable, interesting, and given a style that

the evaluators feel comfortable with. These scores mean little in isolation, however. Averages towards the extremes are of course speaking a clear message, but averages towards the middle gives little information about what worked. This is even more so when we consider that Likert scales are known to have a strong bias towards the centre. And even if the scores average towards “boring” rather than “interesting” (another example from AttrakDiff), there is no way of knowing what it is that makes the product boring, and whether different users are bored by the same aspects of the artefact.

We should also consider what a survey instrument like AttrakDiff actually measures. Hassenzahl’s statistical analysis has shown that it is credible that the measures of *identity, stimulation* and *ergonomics* are separate, and that respondents appear to interpret them in consistent ways. However, we have not found that the authors have analysed whether the semantic differentials actually capture these qualities. To be specific: When users state whether they find a product inventive, creative, bold, captivating, challenging and novel (the positive poles of the seven differentials for "hedonic quality: stimulation"), is there a systematic connection between their answers and the stimulation they experienced?

The two hedonic qualities that are measured with AttrakDiff are identity and stimulation, which are drawn from psychological literature, and believed by the authors to be of major importance when we experience a product as appealing or a joy to use. We may well ask whether we can find other qualities that are equally important, especially for genre design, as different genres are well known to fill different functions. *Stimulation* may be of importance for a pedagogical genre, but for other genres, we might equally well ask whether it inspired feelings of fun, tragedy or suspense — adjectives often used when describing genres in literature. We could apparently make questionnaires measuring fun, tragedy, and suspense using semantic

differentials, although it requires no little work to assure their validity in the rigorous manner Hassenzahl and colleagues have tested AttrakDiff. This work is justified for Hassenzahl as he believes the qualities they measure are universal, rooted in human psychology, and thus applicable to *any* product. Whether we can find such universal qualities for genre design, or indeed if we believe in the possibility of universals, is an open question.

If surveys give little detail, we should realize that the best way of accessing how users experience a new genre is probably to talk to them. Asking evaluators what they thought of the service can be a valuable source of information, but it needs to be carefully monitored. Our experience is that evaluators soon begin to suggest improvements to the service (Nielsen & Loranger, 2006). These are often interesting, and should be collected, but one needs to be careful. We are not always aware of what makes us act in different ways, and what people think they would do in a hypothetical situation with a hypothetical artefact does not necessarily match what they in fact would be doing. More reliable are their reactions to the artefact, both emotionally and intellectually, and this is what the interviewer should be asking for.

Observation, surveys and interviews are established methods in the design sciences. From this little overview, we notice two common traits: First, what these methods do best is to spot failures, or what in design literature is known as "critical incidents." When users aren't able to use the product as intended. When they feel frustrated, or even angry. When they are bored. When they give up, and have to be helped. These are important results both for commercial design and for research. A design that users fail to understand will not do well in a commercial market, and finding such critical incidents early makes it possible to alter the design to avoid that they happen. For research, a failure could be considered a falsified hypothesis, which is the basis for knowledge in many disciplines.

The second common trait is that these methods mainly are based on comparison, whether explicit or implicit. It is difficult for a respondent in a think-aloud study to suggest improvements without pointing to another, existing product. A statistical measure, whether it is response times or average scores in a survey, are only meaningful when compared to the performance of another artefact, whether earlier versions or competing products.

We turn now to the other principle for evaluation in genre design: Evaluating with methods similar to those used to build the theories of communication and genre we built our designs on.

Towards humanist evaluation of computer systems

Humanist research is interpretative, not observational. Its objects are symbolic and meaningful structures made by man, such as writings, music, and visual art, and we who study these structure look for the possible meanings and aesthetic effects they create in readers, listeners, or viewers. Some interpretations aim at finding the exact intention of the author, others look at the meanings that are likely to be found by the audience. Sciences aim to explain and predict natural phenomena by principles that are constant, hence the metaphor of "laws" of nature. Texts and authors, on the other hand, are interpreted. And as each text or each work of art is unique, it cannot be explained by a general law (Gadamer, 2004). In the words of Dilthey, texts are not *explained*, they are *understood*.

The test methods we know from systems design, human-computer interaction, and user experience design all assume a divide between system actions and user interface. All computer systems may be described in this way, and within digital media such as Web or mobile apps, designers routinely describe this divide as "interface" and "content." A simple example of this divide may be a banking system, where the process of moving money from one account to

another is sorted out without user involvement. It just needs “to work,” and the rules by which we judge wheter it is “working” or not are known to everyone. Accessing your funds to pay bills is a matter of moving data in a database.

When it comes to web media, such as a news site, a web TV channel or an online textbook, it is often also viewed as a problem of creating access to a database. Each text is viewed as principally similar, and an interface is made to find the text you wish for. Any text in the system will contain information, but only some text has the specific information the reader wants. Reading is again turned into an access problem: a question of locating an answer within the database of texts. Genre design is different. We are not designing access to generic “content,” we are creating new “content” that is significantly different from earlier “content,” and it is this difference we want to evaluate. There may be initial problems in handling the text with the interface control provided, and these problems may be addressed with evaluation methods from human-computer interaction. But when readers actually get to read the text, how do we evaluate the style, the information, the humour, the drama, the pace — all these things we appreciate when we study genres?

We should remember that texts always have been evaluated, but usually by a small group of experts. In publishing, experienced editors read novels, and coach their authors into making them better according to the editors’ judgement. There is also a different issue in publishing: most publishers receive far more manuscripts than they are willing to publish. As such, the editor can already choose from a large number of prototypes. A similar abundance of manuscripts is found in the film industry, where only a tiny fraction ever gets filmed. In computer science terms, it is an iterative process: A manuscript is selected, re-written, a storyboard is created, before the film is shot, edited and edited again. In all phases there are evaluations in the form of readings and

test runs. When a first edit of the film is (made,) ready, it is showed to a test audience, and their reactions are used to judge whether the edit "works". These are the kinds of evaluation methods we need to develop and make rigorous if we want to advance genre design as an academic practice.

So how can we perform a humanist evaluation? In our project, we will try three ways of doing this: Heuristic evaluation, semantic analysis of user interviews, null hypothesis comparison, and peer review.

User interviews

We have argued that user interviews are the most valuable evaluation method for genre innovation research, and most research projects in the literature have interviewed evaluators after an evaluation session. Krippendorff (2006, p.264;) have suggested a more elaborate method for validation interviews: Early in the design process, we may ask stakeholders what a successful product or artefact should be like. When evaluating the finished artefact, he asks stakeholders to describe it, and compares their descriptions with their earlier accounts of a desirable outcome. Similar descriptions indicate success.

When evaluating of the Rome project, Krippendorff inspired us. We interviewed evaluators after the test and asked them to describe the service in their own words before asking specific questions about the service. These descriptions were later analyzed for semantics to see if they matched our stated research goals, and helped us in answering our research questions. We looked for descriptions and metaphors that gave insights into how the evaluators understood our service, or, as Krippendorff might put it, what *meaning* it had to them. It has to be admitted, though, that the interviews did not yield a lot of insight. Our evaluators did not feel inclined to talk a lot of

their impressions and meaning-making of the service, and it may be that Krippendorff is a bit over-optimistic as to how verbal evaluators usually are.

Another possible and related approach is what Hartson & Pyla (2012) have called 'co-discovery'. This is method where users evaluate a product in pairs, and the test is created in such a way that the evaluators have to talk to each other. Their conversations are recorded, and can later analysed in much the same way as our interview data. This method is difficult to use with an application designed for individuals listening with headphones, however.

It is a question whether we will ever be able to evaluate the finer details in text production with an interview evaluation. In our Rome project we have emphasized tone of voice, reading speed, how to relate historical information about the music with descriptions of its structure and tonality, and how much historical detail about each church is right. We have also discussed whether we should point out certain details within each church, or if we should limit the voice-over to description of the church's totality. We may never be able to find the correct and best solution to these deliberations, our comfort being that the initial evaluastion did not indicate that we have done very badly.

Humanist experiments

Advice and rules for readers are found throughout European history ever since the Greek rhetoric. The basis for this advice must be comparison. Some speeches, tragedies, letters, books, operas or films were clearly better than others, and scholars have studied them in detail to

understand what made them so good.In a second iteration of the project, we introduced this comparative aspect to our project. If the principles for locative audio we have deduced are

robust, audience members should recognize that texts were less good when the principles were *not* followed. So we have authored what could be likened to 'null hypotheses', opposite

examples, texts that deliberately did not follow our own guidelines.

In early April 2013, these texts will be tested in Rome. Each evaluator will be asked to visit two churches and listen to the program, one where the principle is followed, and one were it is not. The table below shows how the test pairs are constructed. We hope this kind of comparison will be a valuable tool for evaluating text production research. Comparative texts can be tailored to answer the researcher's research question, and lead to more informative results than a statement that «evaluators liked what we hoped they would like». A likely outcome of this evaluation is that the evaluators do not experience the differences. Many of the finer details we have worked out (and are rather proud of) may not add much to the users' impressions. If so, we will have to adjust the claims we make for our service. If there is a difference, however, we believe this will be strong support for our design.

Technique	Using	Counter-example
Bringing music back to the original place	Händel in S. Maria Montesantio	The Lateran, any music
Structural similarities between music and architecture	Palestrina in St. Peter's, Medieval chant in S. Cecilia de Trastevere	Il Gesu, San Luigi dei Francesi
Mood fit	Any	Far too fast-spoken example
Music and church from same epoch	Any	Tosca in San Andrea del Valle
Music and commentary	Any	Any

Table 1: For the summative evaluation, we created new texts in pairs to test the strength of some of the guidelines. We tested music with similar mood as the church room versus

contrasting mood, structural similarities in music and architecture versus no similarities, and aura effect versus no mentioning of aura. Evaluators were asked to enter both churches in a pair, and explain which church they liked best and why.

Peer review

It is implicit in what we have written above that while an average user of a location-based system may feel very clearly what works or not for her/him, s/he may not be able to be very specific on why. Most audience members are not authors or analysts, and may never have given the finer details of locative writing much attention. But what about other authors and critics? Peer review is a long tradition in humanist scholarship. Just as scholarly articles are judged by a selection of peer reviewers, books are reviewed by editors before they are published. A feature film regularly goes through stages of review and revision before its theatre release, both in form of manuscript and early edits («rough cuts») of the footage. To let other scholars analyze our texts (or services) should thus be a rather obvious evaluation method, and it has been employed by (2009).

To incorporate this aspect to our production, we will perform at least one evaluation with a scholar with long experience in locative mobile media.

Conclusion

A research agenda for inventing new genres in order to gain new knowledge will need to use existing knowledge of texts, images, and sound for communication as its foundation. If we are serious that we want to do design as research, we must take seriously Hevner, Park, March, and Ram's (2004) argument, that research should aim to achieve new knowledge, to add something

new to what we knew before. It will not suffice to do what is sometimes seen, to build a system and write a paper that describes it, perhaps linking it to some earlier theory. That the author is pleased with his or her system is not a growth in our common knowledge. We need to check our claims, putting them to a test, and consider alternative explanations.

The established methods in design sciences deal with system functions and user interfaces, and are useful in the early phases of most design. But as genres are about symbolic structures that create meaning in an audience, we need to inspect their meaning-making properties when concluding our research. In this paper we have proposed three such methods: Qualitative nterviews that focus on evaluator's meaning-making processes, systematic textual "experiments," painstakingly comparing one aspect after another, and peer review.

We feel certain that design increasingly will be used as a research strategy within media studies, and it is our hope that these methods and other methods that will surface in the future, more advanced and better than these, will help these projects to do better research.

6000 words, including abstract, references, and appendix.

References

Andersson, B.-E., & Nilsson, S.-G. (1964). Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology*, *48*(6), 398. doi:10.1037/h0042025

Bolter, J. D. (2003). Theory and Practice in New Media Studies. In G. Liestøl, A. Morrison, & T. Rasmussen (Eds.), *Digital Media Revisited* (pp. 15-33). Cambridge: MIT Press.

Fagerjord, A. (2011). Between place and interface: Designing situated sound for the iPhone. *Computers and Composition*, *28*, 255-263. doi:10.1016/j.compcom.2011.07.001

Fagerjord, A. (2012). Design som medievitenskapelig metode. *Norsk medietidsskrift*, *19*(3), 198-215.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4). doi:10.1037/h0061470

Gadamer, H.-G. (1960/2004). *Truth and method* (W. Glen-Doepel, J. Weisheimer, & D. G. Marshall, Trans. Second, revised ed.). London: Continuum.

Hartson, R., & Pyla, P. S. (2012). The UX book: Process and guidelines for ensuring a quality user experience. Amsterdam: Morgan Kaufmann.

Hassenzahl. (2000). Hedonic and Ergonomic aspects determine a software's appeal. *CHI Letters*, *2*(1).

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, *13*(4), 481-499.

Hevner, A. R., March, S. R., Park, J., & Ram, S. (2004). Design science in information systems work. *MIS Quarterly*, *26*(1), 75-105.

Krippendorff, K. (2006). *The semantic turn: A new foundation for design*. Boca Raton: Taylor & Francis.

Krug, S. (2010). Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems (Kindle ed.). Berkeley, California: New Riders.

Latour, B., & Woolgar, S. (1986). *Laboratory work: The construction of scientific facts* (Second ed.). Princeton University Press.

Lewis, C. (1982). Using the 'thinking-aloud' method in cognitive interface design. Research report RC 9265. IBM TJ Watson Research Center.

Liestøl, G. (1999). Rhetorics of Hypermedia Design. In *Essays in Rhetorics of Hypermedia Design* (Ph.D. Dissertation ed., p. 265). Oslo: Department of Media and Communication, University of Oslo.

Liestøl, G. (2006). Conducting Genre Convergence For Learning. *Cont. Engineering Education and Lifelong Learning*, *16*(3/4), 255-270.

Løvlie, A. (2009). Textopia: Designing a locative literary reader. *Journal of Location Based Services*, *3*(4), 249-276.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*, 251-266.

Moulthrop, S. (2005). *After the last generation: Rethinking scholarship in the age of serious play*. Proceedings from Digital arts and culture, Copenhagen.

Nielsen, J. (2000). Designing Web Usability: The Practice of Simplicity. Indianapolis: New Rider.

Nielsen, J., & Loranger, H. (2006). *Prioritizing Web Usability*. Berkeley, California: New Rider.

Nyre, L. (2014). Medium design method: Combining media studies with design science to make new media. *The Journal of Media Innovation*, *1*(1), 86-109. Retrieved from https://www.journals.uio.no/index.php/TJMI/article/view/702

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Appendix

Survey questions used in the evaluation of the Church Music in Rome (translated from Norwegian)

1. How did you like this service?

Not at all - not very much - neither much or little - quite a bit - very much

The respondents were asked to score how much they agreed to the following statements on a five-point Likert scale ranging from «totally disagree» to «agree totally».

2. «The talking disturbed me. It would be better to just have the music.»

3. «The music made me experience the church in a different way.»

4. «The comments from the narrator made me experience the music different from how I otherwise would have.»

5. «The music does not fit with the church the way it looks today».

6. «I feel I understand the history of the church now».

7. «When the narrator spoke, I often ‘was lost’ and though about something else.»

8. «I learned something about music history».

9. «It was exciting to be in a place where the music was performed originally».

10. «The whole thing was boring»

11. «I would like to try this service in more of Rome’s churches.»