Realizing the potentiality of metrics in digital preservation

Anssi Jääskeläinen, Mikkeli University of Applied Sciences, Mikkeli, Finland

Abstract

Metrics that are collected by utilizing different methods are the key factors in understanding what is happening under the hood. Currently, the field of digital archiving relies heavily on SLAs (Service Level Agreement) and technical level metrics. While technical metrics are superior in collecting the hard evidence behind any operational unit, it doesn't take account any of the metrics that are relevant when considering the softer side of any technological device, service or software. From the authors' opinion, technical metrics describe merely 1/5 of the whole truth. User and context related metrics aren't utilized and still these have been recognized to be the keys to success in many other fields, in fact these softer metrics are sometimes recognized as a KPIs (Key Performance Indicators). This paper proposes the utilization of the well-known methodologies also in the field of digital archiving in order to gain a good overlook of what is happening behind the scenes.

Starting points

No matter how rewarding or interesting it would be to collect every possible metric from the research point of view and use hours of work to conduct the analysis it's just not possible in the business environment. In the end, in spite of any intangible aspects or technical superiority proven with metrics, it is the incoming money flow that finally convinces the decision makers whether something is good for the business or not.

As a natural continuum for the papers presented in Archiving conferences in 2012 [1] and 2014 [2], this paper continues with the chosen path. The focus is, however pointed more towards the reasons behind utilizing metrics, collecting feedback and gaining deeper understanding of what is happening under the hood as well as what the users are actually doing with the digital archives. By gaining the responses to above mentioned tasks the recognition of the KPIs will be easier.

H. James. Harrington has once said "*Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand it, you can't control it. If you can't control it, you can't improve it"* Therefore, the first step in getting bigger ROIs (Return of Investment) will be to understand the current situation by measuring it with appropriate methods.

The contradiction between money and understanding is a "chicken or the egg" problem. In order to measure and understand the current situation as well as softer values such as usefulness, ease of use or general appealing of the service, utilization of methodologies and tools are needed. This requires time and money. If the treasurer of the company doesn't have a slightest clue about the softer values and the willingness to sacrifice company money for something intangible is zero, nothing happens. The key to get this transmutation ongoing is to realize the benefits of the deeper understanding in the long run and not the expect that the results are immediately visible.

This paper starts by describing the theoretical background behind metrics and SLAs. Secondly, picking the appropriate metrics is considered. Finally, some of the possible methodologies for collecting user and context related metrics are presented. Even though this process might seem straight forward and simple in paper it would have been done ages ago if reality and theory would be easily joined. Simple part is to pick a metric or methodology and just take it into use. Deeper understanding of the method, ability to transform metrics from one field into other and thorough analysis of the harvested results requires experience.

In the previous conferences user experience survey for the users of digital archives [1] as well as archival UI design that was built based on the survey [2] were presented. This paper describes some potential methods that are well known in the user experience community, but haven't been yet recognized in the field of digital preservation. It might seem unrealistic to use methods from the general UX (User eXperience) world in digital preservation, however from the end user perspective there is no difference in using an average website or an archival control UI via web. The same rules and operational modes apply for both cases.

Metrics in generally

A simple and a common way to divide UX metrics into sub categories is by using user, context and system aspects. This classification was originally presented by Hassenzahl and Tractinsky as follows: "*User Experience is a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.)*"[7].

User related metrics are about the user itself and his or her personal preferences, abilities and skills. Context in this case doesn't mean environmental aspects and green values, but the actual surroundings and both physical and organizational conditions where the measured software, service or object is used. For example: You define that your new laptop computer should have a 2k resolution display and a fast display adapter, but you don't mention anything the usage environment. What will happen if the office where the laptop is used doesn't have window blinds and the sun is shining directly to your screen? You either complain to the computer vendor that the expensive computer is useless or you start to struggle with the space administrators in order to get window blinds. Either way this little incident would have been avoided if a context related metric called luminous flux would have been taken into consideration earlier.

In the digital archiving field, the system metrics consists of up- and down time, latency, IOPs, system load, etc. Technical metrics are the ones that are generally defined in SLAs or the service level is somehow calculated based on the low level technical metrics [12]. Another side of the technical metrics is the agreed system requirements which completeness is checked during the acceptance testing.

For the service that is already running, the technical metrics are also the simplest one to handle and the collection is relatively simple to automatize. However, possibility to automatize collection is not necessarily a good thing. The following listing shows some very commonly collected technical metrics which weight, from authors' opinion, is virtually close to a zero in spite of the case.

- **Access counter:** While this easily accessible integer might be a way for tracking the efficiency of advertisements it doesn't actually tell anything about the visitors experience with the site or service. Furthermore, nowadays even if the access counter records IP address and system identifiers of every visit it is still somewhat difficult to say was there a human operator or was the site accessed by a bot.
    - Instead you could measure the success rate of every unique access. For example if you have a site that sells something you could measure the percentage of the successful sales of unique visits.
- **Time on site:** In generally more time spend on the site or service is considered as a good thing. But, it might also reflect usability or other problems which force visitors to stay longer that they would like.
    - Measure the time between certain key actions that the visitors are likely to take. This way you are able to find and identify the bottle necks of your service design.
- **Email subscribers:** One of the very common metrics especially if you are doing email marketing. Still just the size of the list doesn't tell anything important
    - Measure the email opens, clicks on the email that led to your site and lastly measure the amount of visits that originated from the email and led to purchase. This way you have the efficiency of your email list, not just the size.

SLA guarantees everything

Unfortunately, this is a very common misunderstanding that the client generally counts on. Hate to break it, but technically superior system that has 99.999% SLA guaranteed up time and latency time less than 1ms can be a real pain in the bum for the users of the system. Reason can be for example an inconsistent UI, bad search and index features or even the above mentioned display in too luminous conditions.

Some examples of such a technically splendid systems and their usage experience can be found in the responses and statements that the respondents gave in 2012 in a survey about digital archive user experience [1]. Survey was conducted in Finnish so it is somewhat difficult to translate the statements, but those were not too flattering. One respondent for example stated that a particular "unfinished" information management system should not even have been released yet. This particular IMS system was done by a big software vendor and had most likely gone through an extensive acceptance and technical tests before the launch. This said, technical perfectionism and promised service levels only ensure one of the five core user experience attributes that every product, service or application should have, performance. The rest of the core attributes; usability, appeal, accessibility and user assistance / help are either overlooked or completely forgotten.

It would be simple to support the automatically collected technical metrics by e.g. collecting regular customer feedback or by conducting some qualitative user studies. However, it is the SLA that state is this extra work relevant to the business or not. If the softer side of the metrics isn't mentioned in the contract and there is no competition in the field, what is the point of doing extra work from the business point of view?

From the plain theoretical point of view there is a lot more that could, and should be done in order to gain a good overlook of the customer relationship. When it comes to the user and context related metrics, the author hasn't encountered a single SLA which would have defined either of these even slightly. There is a good reason for that. It isn't simple to gather, analyze or especially understand these very subjective metrics.

The knowledge behind these subjective context- and user related aspects is the base for understanding and control, which on the other hand forms the basis for improvement. The author however agrees that the place for these softer values is not in the SLA. Instead it should be included in the general agreement that also the non technical metrics should be measured and considered.

The good and the bad metrics

While collecting metrics, it might sound like a good idea to gather every possible metric that the utilized analytic tool offers. For example, Google analytics account can be set up in few minutes and the received JavaScript can be placed on the global site element. After this simple action, an access to hundreds of reports and more metrics that can ever even be imagined is available [4]. But how to know which ones to use to gain the best outcome? If and when you cannot decide, a backup solution might be to use them all just in case. This "just in case mentality" is however a very bad choice in the case of selecting metrics. Stored raw data that was supposed to be the savior the business after analyzed properly becomes a burden, since the amount of data soon becomes too large to handle. As a result of overwhelming amount of information, nobody cares, or doesn't have time to handle it. Key is to thoroughly analyze the business environment and according to analysis pick the appropriate metrics, KPIs for the purpose.

Picking appropriate metrics and methods

There are many ways to select the utilized metrics. The purpose of this paper is not to introduce these ways, just to show that methods and metrics from the UX world are also usable in the field of digital archives. However, a selection method called the So What Test is presented thus from authors' opinion it is one of the easiest one to conduct and it doesn't require background knowledge of metrics or analysis.

The So what test that was originally introduced by Kaushik[3] is a three layered test method for selecting metrics. The test is based on asking so what question for every measured metric for three times. If this question cannot be answered purposefully in each question round, the metric most likely won't be necessary. Following two examples will demonstrate the utilization of the so what test.

1. Our digital archive system email news letter have had 30% increase in subscribers since 2014. *So what?*
2. People are obviously more interested in archiving and they are also more aware of our services. *So what?*
3. Maybe they will now start buying our digital archive service. *So what?*
    - If the answer to the last *so what* question is something like “Isn't this nice to know.. or I don't know or something similar that cannot be validated or ensured this particular metric might not be the best for you.

1. Our up time has increased to 99.7% from 98.9% after we switch our hardware vendor. *So what?*
2. Their products are obviously more reliable than the previous ones. *So what?*
3. We should replace all of our hardware with hardware from the new vendor since now we fully meet the SLA requirements. *OK.*

From these two examples, the first one would lead to either abandoning the metric or modifying it to measure e.g. the access to the site via posted email attribute. On the other hand, the second example would be accepted since it leads to a certain defined answers.

When speaking about picking the appropriate metrics in general, only the things that can be changed and you or your company is willing to change, should be measured. If you cannot take the necessary actions suggested by analysis of the metrics, why are you even collecting the metrics? Don't for example measure the efficiency of the metadata input form if there is no intention to change the form or some legislative thing prevents from changing it. Finally, remember that in the case of metrics quantity won't be an alternative to quality.

Possibilities of the metrics

The digital archive world relies on technical metrics, but the user experience world is full of potentially usable methods that produce user- and maybe even context related metrics. Measuring the subjective user experience attributes is for example an everyday action for example in the area of software development. Nearly hundred different methods, tools and theories exist about how even the intangible and subjective attributes can be acquired and analyzed.

The archival field has been different for multiple reasons. The first reason is the domination of very minor amount of commercial vendors. There hasn't been enough competition to rationalize the extra efforts for measuring subjective attributes. The second reason is the ignorance of the users that still live according to the rules from the world of paper [2]. In other words, users don't even realize that with a modern technology things can be different. Finally the general development mentality “by developers for developers” is still very dominating in many technological fields where the paying client is absolutely the last one who sees the product.

From the authors' opinion, which is backed up by Jokela [11], for instance, usability and user experience work should start as early as in the requirement definition phase. Currently, the requirement definition might state that "the system must be easy to use" and that "the system must make it possible to add and alter the metadata of an object". The metadata part can be validated, but from the requirement perspective it is the same if this metadata input can be done with one mouse click and one input field or with 30 mouse clicks, five different menus and 20 input fields. The easy to use statement is impossible to validate if no guideline for “easy to use” is given. Every software- or system vendor can just announce that their system is easy to use according to their own tests, nothing else is needed unless “easy to use” is defined in a measurable way.

To start with

This chapter introduces some of the most appealing subjective tools that can be utilized to gather the user and context related aspects. In the modern digital world, many of the products to be evaluated are already on the market. Therefore, the focus is on the evaluation methods that are suitable for the products on the market. Luckily, there are still circa 70 possible methods from which to choose from.

Affect grid

Simple evaluation method where user marks his/her current emotional state in to a 9x9 grid, which is presented in Figure 1. Method is originally designed to be a quick way to assess users' emotions in pleasure-displeasure and arousal-sleepiness scales. Since this method is a single-item scale it can be used rapidly and repeatedly. Even though this method is not so reliable as multiple-item questionnaires there still is a strong evidence of its power [6]. This method has been used for example in measuring the emotions in software requirements engineering [10]. Could easily be used also in the archival field to find out the general user related opinions about the archival system. One aspect that this method does not take into consideration is time, but that can be achieved by running this test for example right after the first contact and after several months of usage.

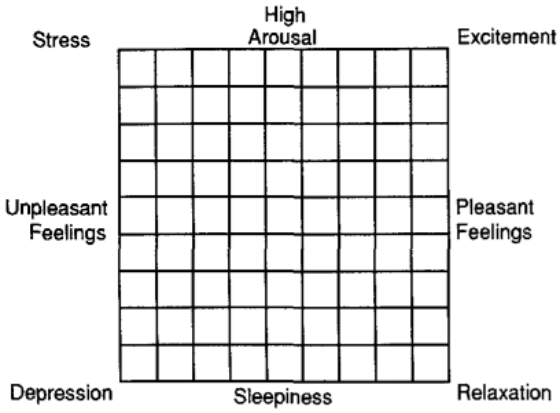

Figure 1. *Original Affect Grid by Russell, Weiss and Mendelsohn [6]*

Co-discovery

Two participants who are preferably familiar with each other explore the product together and conduct some predefined test cases. Participants who are familiar with each other are more likely to speak freely about the task and the possibility for uncomfortable feeling is lower than with participants who are strangers to each other. While conducting the predefined test cases, participants are encouraged to discuss the subject of the test as well as describe their thoughts. Video / voice recording is used to capture the discussions [9] as well as either direct or indirect observation by a usability professional or psychologist.

Co-discovery method is a relatively simple to set up, but it requires twice as much participants than for example the self-discovery methods such as thinking-aloud. On the other hand co-discovery produces better understanding of the relationships between the actions taken during the tests and the goals of the test. Further on, this tends to lead to more affective test results. Co-discovery is a very usable method but analysis take time and require knowledge of human behavior, e.g. basics of psychology or usability.

Day reconstruction method

The Day Reconstruction methods Or DRM tries to assess how participants spend their time during the test session. This method is most commonly used to find out behavioral patterns of participant during a longer period of time for example in eating disorder cases. However, this method is also very usable when resolving patterns how people behave while they are conducting some predefined tasks. Participants systematically reconstruct their activities and experiences of the preceding day with procedures designed to reduce recall biases [5]. Experience with the product is captured for example in a manner of three of the most meaningful (either good or bad) with detailed descriptions of the situation and context. Simple to use but relatively laborious to analyze properly.

Valence method

User experience is defined by Hassenzahl as an evaluative feeling during the usage [8]. The Valence methods is based on capturing this evaluative feeling during the usage by pressing a dedicated physical button for a feeling. Normally, this button is either a green plus or a red minus. Button presses are registered as valence markers with time stamps on recorded video. After the first phase a retrospective interview will happen where test users watch the recording and comment on what they were experiencing at each valence marker. Then the interviewer asks which product feature, element or function element caused the button press and in co-operation with the user tries to identify the underlying need. Although this method is simple to set up, the analysis phase requires an understanding of emotion and motivation based design.

Conclusions

There are many suitable methods, which should be carefully chosen with a trained professional in order to gain the best possible reflection from the users of the system. It needs to be kept in mind that most of the available methods only produce raw information, which needs to be analyzed by a trained professional. Another alternative is to develop an automated or semi-automated algorithm for the analysis with the aid of a professional.

One of the worst mistakes that can be done is to allow the designer or programmer of the system to conduct the analysis. While these persons are good at what they are hired to do, they are most likely not capable of conducting thorough analysis of subjective information.

In this paper, the potentiality of different metrics and methodologies were presented. The world is full of metrics and this overwhelming amount of possibilities can be a blessing or a curse. It is a blessing for those who are either eager to learn how to utilize those or the ones how are currently familiar with different methodologies. However for the rest of possible users, roughly 99.9999% it is a confusing mess. Some ways, such as the So What Method, to make the mess somewhat more tolerable, were presented. Finally, this paper also presented some simple methods that can be utilized even if the know-how is close to zero. In the case of metrics, quality over quantity and not vice versa.

References

[1] A. Jääskeläinen, Rationalizing the concept of user experience in digital preservation, Proc. Archiving 2012 (2012)

[2] A. Jääskeläinen, T. Vuorikari, Smoothing away the relic of the past: Case archival control UI, Proc. Archiving 2014 (2014)

[3] A. Kaushik, Web Analytics: An Hour a Day (Wiley publishing, Indianapolis, 2007)

[4] B. Clifton, Advanced Web Metrics with Google Analytics (Wiley&Sons , Indianapolis, 2012)

[5] D. Kahneman, A. B. Krueger, D. Schkade, N. Schwarz, A. Stone. The Day Reconstruction Method (DRM): Instrument Documentation, Online avail. http://sitemaker.umich.edu/norbert.schwarz/files/drm_documentation_july_2004.pdf

[6] J. Russell, A. Weiss. G.A. Mendelsohn. "Affect grid: a single-item scale of pleasure and arousal". Jour. Of personality and social psychology.57, 3, (1988).

[7] M. Hassenzahl, N. Tractinsky. "User experience – a research agenda" Jour of Behaviour & Information Technology, 25, 2, (2006)

[8] M. Hassenzahl. User Experience (UX): towards an experiential perspective on product quality, Proc IHM'08

[9] P. W. Jordan, Designing Pleasurable Product (Taylor & Francis, London, 2000)

[10] R. Colomo-Palacios, C. Casado-Lumbreras, P. Soto-Acosta, A. Garcia-Crespo. "Using the Affect Grid to Measure Emotions in Software Requirements Engineering" Jour. of univ. computer. sci. 17, 9, (2011)

[11] T. Jokela, Determining usability requirements into a call-for-tenders: a case study on the development of a healthcare system, Proc NordiCHI '10 pg.256, (2010)

[12] V.C Emeakaroha, L. Brandic, M. Maurer, S. Dustdar. Low level Metrics to High level SLAs - LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments (HPCS 2010, Caen France) 48-54.

Author Biography

Anssi Jääskeläinen has an IT MSc. (2005) from Lappeenranta University of Technology and a PhD (2011) about considering user experience in software development. He has an extensive knowledge of user experience and usability. His current interests are in user experience, Java EE and game engines. He currently works as a head lecturer at Mikkeli University of Applied Sciences