

Ganzheitliches UX-Measurement im Großunternehmen

Strategien, Prozesse und Tools zur kontinuierlichen Verbesserung der internen IT-Lösungen

Marian Gunkel
IT Services
SAP SE
Berlin, Deutschland
m.gunkel@sap.com

Maike Vocke
IT Services
SAP SE
Berlin, Deutschland
maike.vocke@sap.com

ABSTRACT

Eine UX-Strategie gehört heute zum guten Ton – doch wie misst man ihren Fortschritt? Für die nutzerzentrierte Weiterentwicklung interner Prozesse und Anwendungen nutzen wir seit 2012 ein kontinuierliches und standardisiertes Messverfahren anhand von Online-Surveys. Mit Hilfe von quantitativen und qualitativen Daten messen und erklären wir die Nutzerzufriedenheit. Multiple Vergleiche sind durch den standardisierten Ansatz möglich und ermöglichen weitergehende Analysen und evidenzbasierte strategische Entscheidungen. Produkt-Teams messen ihre Zielerreichung anhand der KPIs und steigern so den UX-Reifegrad des Unternehmens. Ein produktzentriertes Dashboard verknüpft die aus den Online-Surveys gewonnenen Nutzeroaussagen mit Verhaltensdaten und Accessibility-Kennzahlen zu einem ganzheitlichen Bild. In kontinuierlichen Messungen entsteht so über die Zeit ein ganzheitliches Bild aus Produktverbesserungen und daraus folgenden Änderungen im Verhalten und Erleben der Nutzer.

KEYWORDS

UX Strategie, ResearchOps, Measurement, UX Maturity, Fragebögen, Methodenkombination, Kennzahlen

ACM Reference format:

Marian Gunkel, Maike Vocke. 2019. Ganzheitliches UX-Measurement im Großunternehmen: Strategien, Prozesse und Tools zur kontinuierlichen Verbesserung der internen IT-Lösungen. In *Mensch und Computer 2019 – Workshopband*, Bonn: Gesellschaft für Informatik e.V., <https://doi.org/10.18420/muc2019-ws-653>

1 Einführung

Die interne IT der SAP verantwortet die Arbeitswerkzeuge von etwa 100.000 Angestellten und freien Mitarbeitern weltweit,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MuC'19 Workshops, Hamburg, Deutschland

© Proceedings of the Mensch und Computer 2019 Workshop on Konstruktion und praktischer Einsatz von User Experience Fragebögen. Copyright held by the owner/author(s).

<https://doi.org/10.18420/muc2019-ws-653>

darunter SAP-Standardprodukte, aber auch viele Eigenentwicklungen. SAP's ehemaliger Chief Information Officer (CIO) Thomas Saueressig hat wiederholt glückliche Mitarbeiter als die entscheidende Voraussetzung für ausgezeichnete Produkte und Lösungen und damit glückliche Kunden bezeichnet und somit „user-centric IT“ als wichtiges Ziel etabliert [1]. Die allgemeine Mitarbeiter-Zufriedenheit wird in jährlichen Befragungen erfasst, wir konzentrieren uns in diesem Beitrag auf die spezifischeren Aspekte der Arbeitsmittel und -umgebungen.

Doch was genau ist nutzerzentrierte IT, wie lässt sich deren Reifegrad messen und, am wichtigsten: wie lässt sich diese nachhaltig als Mindset im Unternehmen etablieren?

In unserem Beitrag soll es hauptsächlich um die ersten Schritte gehen: hier sind vor allem arbeitsmittel-spezifische Messungen wichtig, um sowohl Transparenz über den aktuellen Status Quo herzustellen als auch konkrete Verbesserungen der Produktivität und Zufriedenheit der Nutzer zu ermöglichen.

2 Strategischer Ansatz

Für uns war zunächst wichtig, verlässliche Daten zu gewinnen, anschließend die Skalierbarkeit der Messungen zu gewährleisten (Stichwort ResearchOps) und diese schließlich in einen ganzheitlichen Feedback-Regelkreis zu überführen. Unterschiedliche Informationsbedürfnisse für diverse Zielgruppen (vom Produkt-Owner und UX-Strategieberater über verschiedene Management-Ebenen bis hin zum CIO) sollten ebenfalls berücksichtigt werden.

2.1 Valide Messinstrumente und Daten

Die Basis unseres Prozesses bilden Online-Surveys, die anwendungsspezifisch und wiederholt die Nutzungszufriedenheit von Endanwendern messen.

Im Jahre 2012 haben wir uns bewusst für den SUS (System Usability Scale) Survey [2] entschieden. SUS ist ein valides und reliables Messinstrument [3], in vielen evaluierten Sprachversionen verfügbar, sowie sehr schnell durchzuführen und einfach auszuwerten.

Wir erweitern den SUS-Fragebogen sowohl durch die Abfrage von demografischen Daten (z.B. Land, Rolle oder Nutzungshäufigkeit) als auch durch den Net Promoter Score (NPS). Darüber hinaus

erfragen wir per Freitext die Stärken und Schwächen der Software aus Nutzersicht. So können wir sowohl die Höhe der Usability-Zufriedenheit, den Einfluss von demografischen Faktoren als auch diverse Ursachen für Nutzungsprobleme erfassen.

2.1.1 Quantitative Daten. Der quantitative Teil unserer Umfragen enthält drei Elemente: den SUS-Fragebogen, den Net Promoter Score (NPS) und demografische Daten. Der Fragebogen, aus dem schließlich der SUS-Score ermittelt wird, besteht aus zehn Likert-Skala-Aussagen, welche den Grad der Usability aus verschiedenen Perspektiven adressieren (siehe Abbildung 1). Der SUS Fragebogen als ein standardisiertes Messverfahren ermöglicht es uns, Umfrageergebnisse toolübergreifend miteinander zu vergleichen.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I like using frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I find unnecessarily complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that is easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that I would need the support of a technical person to use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I find the different functions of well integrated together.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
There is too much inconsistency in	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that most people can learn to use very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I find very cumbersome to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel very confident using	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I needed to learn a lot of things before I could get going with	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Abbildung 1: SUS Fragebogen

SUS Scores korrelieren in der Regel stark mit Net Promoter Scores, welche die Weiterempfehlungsrate einer Anwendung anzeigen: Denn je zufriedener die Nutzer mit einem Tool sind, desto größer ist die Wahrscheinlichkeit, dass sie ebendieses Tool einem Kollegen weiterempfehlen werden.

Ein ähnlicher Zusammenhang lässt sich auch zwischen dem SUS Score und den demographischen Daten feststellen, die wir in unseren Surveys erheben. Hier sei beispielhaft vor allem die Frage nach der Nutzungshäufigkeit genannt. Je häufiger eine Anwendung verwendet wird, desto zufriedener sind die Nutzer mit ihr. Dies ist auf Lerneffekte zurückzuführen, die sich bei einer häufigen Nutzung schneller einstellen (siehe auch [3], Seite 95ff). Der SUS Fragebogen wird also sowohl durch den Net Promoter Score als auch durch die Abfrage von demografischen Daten sinnvoll ergänzt.

2.1.2 Qualitative Daten. Der qualitative Teil der Survey besteht aus drei Fragen, bei denen die Nutzer Vorschläge für Verbesserungen machen und angeben können, welche Aspekte der entsprechenden Anwendung sie am meisten und am wenigsten mögen. Da diese drei Fragen durch ein offenes Textfeld beantwortet werden, gehen wir davon aus, dass nur die Aspekte genannt werden, die für die Nutzer von Relevanz sind. Auch anhand der Häufigkeit (vgl. [4]) und der Tonalität in der bestimmte Themen angesprochen werden, lässt sich die Relevanz eines Themas ausmachen.

Die Stärken und Schwächen der jeweiligen Software werten wir mit einem selbstentwickelten Kategoriensystem aus. Dieses beinhaltet diverse „klassische“ Usability-Aspekte (u.a. visuelles Design, Terminologie, Kontrollelemente, „Microcopy“), Aufgabenangemessenheit (z.B. Passung zum Geschäftsprozess, Service Design, persönlicher Mehrwert, Kommunikation, Integration mit anderen Tools), Inhalt (z.B. Datenqualität, Informationsarchitektur) und Technologie (z.B. Bugs, Performance, Verfügbarkeit). Durch die so zugeordneten Kommentare wird die Relevanz der jeweiligen Usability-Kategorien wieder zählbar. Weiterhin ist es möglich aus den Kategorien und dem Nutzerfeedback konkrete Handlungsanweisungen abzuleiten und zu priorisieren.

2.1.3 Frequenz. Aus Kapazitätsgründen haben wir zunächst pro Software eine Messung pro Jahr durchgeführt. Versuchsweise im Quartalsabstand durchgeführte Messungen des Tools „Corporate Search“ brachten statt der erhofften stärkeren Aufklärung eher Frust: die Kennzahlen SUS und NPS blieben trotz größerer Produktänderungen zunächst auf ähnlichem Niveau und stiegen erst mehrere Monate später. Die Produkt-Teams hatten eine schnellere Änderung der Nutzungszufriedenheit erwartet, die Nutzer waren wiederum durch die „gefühlte“ häufigen Befragungen genervt. Anscheinend beeinflussen Produktverbesserungen erst nach häufiger Interaktion die subjektive Qualität der Software und das anschließende Urteil der Teilnehmer verbessert sich nur langsam. Einzelne negative Nutzungserfahrungen werden deutlich besser erinnert als typische positive Nutzungserfahrungen und färben die Beurteilung auch Monate oder sogar Jahre später.

Trotzdem planen wir häufigere Messungen mit kleineren Teilnehmergruppen, da wir so per Frühwarnsystem von größeren Problemen erfahren können.

2.1.4 Rekrutierung. Wir verwenden zwei unterschiedliche Rekrutierungsmethoden, die jeweils Vor- und Nachteile mit sich bringen. Für Software, die typischerweise unternehmensweit genutzt wird (z.B. Intranet, Urlaubsantrag, Adressbuch), nutzen wir ein nach Unternehmensbereich und Region geschichtetes zufälliges Sample von 500-1000 Personen, an das anschließend per Email Einladungen verschickt werden. Dieses Schrottschuss-Prinzip erreicht auch Selten- oder Nicht-Nutzer einer Software und deckt somit auch interessante und wichtige Perspektiven dieser Nutzergruppen ab. Es führt aber auch häufig zu indifferenten oder im schlimmsten Falle verfärbten Daten, da die

Selten-Nutzer überdurchschnittlich häufig negative Nutzungsaspekte als Basis ihrer Beurteilung zu nutzen scheinen. Wir glauben, dass das „Technology Acceptance Model 3“ [5] dieses Phänomen gut erklärt: dessen Faktoren „Perceived Usefulness“ und „Perceived Ease of Use“ sind für Selten-Nutzer durch eher rudimentäres Erfahrungswissen und schwache mentale Modelle beeinflusst. Für spezialisierte Software (z.B. für Entwickler, Controller, Vertriebler) laden wir direkt im Tool (z.B. per Pop-up, Announcement, floating button) zum Survey ein, erreichen so aber natürlich nur aktive Nutzer. Deren Urteil erweist sich jedoch häufig als realistischer und basiert eher auf der aktuellen Nutzung, weniger auf früheren Erfahrungen. Unsere Planung sieht vor, beide Rekrutierungsmöglichkeiten stärker miteinander zu verzahnen, um die jeweiligen Nachteile auszugleichen.

2.1.5 Datenschutz. Seit Beginn unserer Messungen im Jahre 2012 legen wir Wert auf absolute Anonymität der Teilnehmer. Außer dem Zeitstempel der Antwort und den selbstgewählten demografischen Angaben werden keine persönlichen Daten gespeichert. Die Versendung der Einladungs-E-mails erfolgt ohne Tracking, auch der Aufruf des Surveys direkt aus dem Tool heraus wird nicht registriert.

2.1.6 Benchmark. Für SUS und NPS haben wir seit dem Jahre 2012 einen internen Benchmark aus ca. 15.000 Messungen aufgebaut. Diesen erweitern wir beständig durch Ergebnisse aktuellerer Surveys.

2.2 Vergleiche und Zielsetzungen: Mittel zur strategischen Verbesserung der Nutzerzentrierung

Durch die Standardisierung und Wiederholung der Messungen können wir auf quantitativer und qualitativer Ebene erstens Vergleiche über die Zeit per Anwendung und zweitens zwischen Produkten anstellen. Der zeitliche Trend der Usability einer Anwendung bildet den Schwerpunkt unserer Arbeit. Häufig wird jedoch auch die Fragestellung beantwortet, ob, z.B. bei der Reisekostenabrechnung, ein Vorgängerprodukt durch den Nachfolger verbessert wurde.

Die Ergebnisse der Surveys werden auf Produkt-Ebene genutzt, um einerseits Entwicklungsaktivitäten zu priorisieren, andererseits aber auch zu de-investieren (Produkt ist „gut genug“ und Ressourcen werden auf problematischere Produkte verschoben).

Mittlerweile werden von vielen Teams und Managern die KPIs (SUS und NPS) genutzt, um jährliche Team- und Manager-Ziele zu vereinbaren und zu messen. Je nach Status Quo liegt das Ziel entweder in einer deutlichen Steigerung der Kennzahlen oder auch, bei einem hohen Level der Kennzahlen, im Halten dieses Niveaus. Letzteres Ziel ist primär bei geplanten neuen Funktionalitäten und höherer Komplexität des Produktes wichtig.

Durch die Verankerung in den individuellen und Organisationszielen wird UX zu einem Teil des Mindsets in der IT-Organisation.

3 Ausblick: ein ganzheitliches UX-Dashboard

Doch die allein auf Nutzeraussagen basierenden Einsichten reichen uns nicht. Wir wollen auch das Verhalten über die Zeit sowie den Grad der Accessibility erfassen und diese unterschiedlichen Aspekte an einem Ort zusammenführen.

3.1 Nutzungsverhalten

Da Nutzungsverhalten hoch produktspezifisch ist, haben wir unseren standardisierten Ansatz in diesem Bereich verlassen und pro Anwendung Verhaltensmuster identifiziert, die 1. einen guten Einblick in das Verhalten geben, 2. auch in Logfiles getrackt werden können und 3. für das Produkt-Team verständlich und umsetzbar sind.

Ein Beispiel anhand der Corporate Search (Suche über eine Vielzahl von unterschiedlichen internen Wissensplattformen) ist die „Anzahl der Änderungen der Suchbegriffe während einer Session“: je höher dieser Wert, desto schlechter ist wahrscheinlich die subjektiv wahrgenommene Qualität der Suchergebnisse.

Im Gegensatz zu den standardisierten Fragebögen-Kennzahlen betrachten wir die Verhaltensmuster vorerst als „schwache Signale“, da wir deren Fehlerrate nicht genau bestimmen können. Erst wenn mehrere dieser Signale deutlich vom Mittelwert abweichen, nehmen wir eine problematische Nutzungserfahrung an und betreiben Ursachenforschung.

3.2 Accessibility

Die Barrierefreiheit einer Anwendung messen wir durch einen auf Grundlage des SAP-Accessibility-Standards entwickelten Fragebogen, der von betroffenen Kollegen ausgefüllt wird. 25 Fragen zu unterschiedlichen Aspekten der Barrierefreiheit (z.B. Attribute von UI-Elementen, visuelle Darstellung, Interaktion unter zeitlichen Aspekten, Orientierung und Navigation innerhalb der Anwendung) werden auf einer dreistufigen Skala beantwortet und ergeben einen prozentualen Score.

3.3 Dashboard

Die Daten werden in ein produkt-zentriertes Dashboard überführt, um alle relevanten Informationen an einem Ort zu sammeln. Dort können POs, aber auch UX-Berater, Veränderungen im Nutzererleben und -verhalten in Verbindung mit vorherigen Produkt-Releases bringen und objektiv erfassen, ob bestimmte Produktänderungen die UX maßgeblich beeinflussen.

Aktuell befindet sich das Dashboard im Prototypen-Status, da noch viele Fragen geklärt werden müssen: wie ist die beste Darstellung für welche Nutzergruppe? Wer darf Zugriff auf die teilweise sensiblen Daten haben? Ist ein manuelles Update ausreichend oder wollen wir so weit wie möglich automatisieren?

3.4 Feedback-Loop

Zukünftig soll ein Kreislauf entstehen: Nutzerfeedback führt zu Produktänderungen, diese beeinflussen Verhalten und Erleben der Nutzer, welche wiederum gemessen werden und als Grundlage für potentielle weitere Produktänderungen dienen.

4 Die Organisationsentwickler-Brille

Management-Unterstützung ist die wichtigste Komponente für die Durchsetzung des Ansatzes in der gesamten Organisation. Viele Teams konnten wir inhaltlich überzeugen und mit deren Hilfe unseren Prozess verfeinern. Erst durch den Ansatz des CIO ist jedoch auch top-down das Thema UX in der IT verankert worden, unter anderem durch entsprechende Ziele für das Senior Management.

Wir befinden uns mitten in einer Kulturänderung, die wir nach Kräften selbst beeinflussen und für die wir sowohl Daten liefern als auch Vorbild für Mindset-Änderungen sind.

ACKNOWLEDGMENTS

Dank gebührt unserem Manager Ingo Widmann, der uns stets viel Freiraum für Experimente ermöglicht. Darüber hinaus danken wir Bernard Rummel und Martin Schrepp für die Einbindung in den Workshop. Vor allem aber sind wir unseren vielen tausend Kollegen weltweit dankbar, die uns als Survey-Teilnehmer einen so reichhaltigen Datenschatz ermöglicht haben.

LITERATUR

- [1] Saueressig, Thomas. 2016. SAP-Digital-Transformation und benutzerzentrierte IT-Vision. Vortrag auf SAP-interner Versammlung: Coffee Corner Session, SAP SE Berlin.
- [2] Brooke, John. 1996. SUS: a "quick and dirty" usability scale. In: P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Ed.): Usability Evaluation in Industry. Taylor and Francis, London, UK.
- [3] Sauro, Jeff. 2011. A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Measuring Usability LLC, Denver, CO, USA.
- [4] Brunner, E.J. & Tschacher, W. 1995. Quantifizierende Inhaltsanalyse. In König, E. & Zedler, P. (Hrsg.): Bilanz qualitativer Forschung, Bd.2: Methoden, Weinheim: Deutscher Studienverlag.
- [5] Venkatesh, Viswanath & Bala, Hillol. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. Decision Sciences - DECISION SCI. 39. 273-315. 10.1111/j.1540-5915.2008.00192.x.