Thesis

Exploring Big Data Visualization for the Digital Publishing Industry

Improving Insight for Publishers using Visualizations on imgZine's Real Time Social Magazines Platform

Final Thesis for the
Master's Program in
Business Information Technology
at imgZine,
Amsterdam, the Netherlands

B.J. van der Wees
University of Twente
Enschede, the Netherlands
b.j.vanderwees@alumnus.utwente.nl

imgZine

UNIVERSITY OF TWENTE.

Example of cover view

in order to obtain the
degree of Master of Science
at the University of Twente,
by authority of the Exam Committee BIT,
to be publicly defended
on Friday the 13th of December, 2013

by

B.J. van der Wees
born February 18, 1988
in Utrecht, the Netherlands

Example of index view

Student Information

B.J. van der Wees BSc
b.j.vanderwees@alumnus.utwente.nl

Supervisory Committee

First supervisor
Dr.ir. Hans Moonen
Assistant Professor
Faculty of Management and Governance
Industrial Engineering and Business Information Systems (IEBIS)
hans.moonen@utwente.nl

Second supervisor
Dr.ir. Maurice van Keulen
Associate Professor Data Management Technology
Faculty of Electrical Engineering
Mathematics and Computer Science (EEMCS)
m.vankeulen@utwente.nl

External supervisor
Marijn Deurloo MBA
CEO & founder imgZine
marijn@imgzine.com

External Organization

imgZine B.V
Basisweg 52D
1043 AP Amsterdam
The Netherlands
+31 20 411 18 38
info@imgzine.com

iPad 11:06 38%

Dienstverlening

Samen sterker

22 november 2013

Blog van Nynke Struik
Een half jaar geleden, bij de lancering van de Rabo Kennis App, ben ik begonnen met bloggen. U heeft titels voorbij zien komen als 'meer galini', 'voel het verschil' en 'driehoeksrelatie'. Losse flodders? Zomaar gedachtenspinsels? Zo lijkt het misschien wel. Toch zit er een rode draad in.

Bij de lancering van de Rabo Kennis App schreef ik mee aan de publicatie 'Kennis delen, laat kennis groeien' (PDF, 128 kB). Hierin schrijven we dat kennis essentieel is voor ondernemers. In een omgeving die snel verandert en waarin technologische ontwikkelingen elkaar razendsnel opvolgen, zijn ondernemers genoodzaakt om zich voortdurend aan te passen aan ontwikkelingen in de markt. Kennis ove uw sector en de ontwikkelingen in de financiële markten, biedt uw bedrijf een voorsprong en de mogelijkheid u te onderscheiden van de concurren…

Cijfers & Trends

Contact

Internet-of-things ideale biotoop voor worm

1 hour ago

Een nieuw ontdekt stukje malware, Linux.Darlloz genaamd, dreigt zich te nestelen in het ingebedd…

Regeling cofinanciering sectorplannen aangepast

An 1 hour ago

De Regeling cofinanciering sectorplannen is aangepast naar aanleiding van de begr…

Ook zonder iPhone betaal je Apple voor de iPhone

2 uur geleden

De voorwaarden die Apple weet te bedingen bij mobiele-telefoniebedrijven drijven de kosten voor hun…

Nederland geprezen om 'nood roaming'

2 uur geleden

Het systeem waarbij mobiele telecomaanbieders elkaars netwerk kunnen gebruiken in geval van grote calamiteiten heeft internationaal waarde…

Thuiskopieheffing op helling door 'Acer-vonnis'

2 uur geleden

Een uitspraak van de rechtbank in Den Haag, vorige week, kan grote gevolgen hebben voor de…

Windows 7 groeit harder dan Windows 8

3 uur geleden

Microsoft moet met lede ogen toezien hoe een 3 jaar oud besturingssysteem sneller marktaandeel wint dan het vorig jaar geïntro…

Rabobank Kennis

1 / 13

Login met LinkedIn

Example of magazine view

Preface

In front of you, there is either a paper, or in case that you are lucky a digital version of this Master's thesis.

Where most people are able to read from paper, digital solutions are not always as accessible and usable as they should be. I believe everybody should have access to new technologies in an easy way. And therefore both my bachelor and master thesis present a search for new and easier IT solutions. Although current solutions might exists, there is always room for improvement. After my Bachelor's explorative study on In-Train Crowdsourcing at the Dutch Railways (van der Wees & Moonen, 2011) I decided to do another explorative study, this time in the field of Big Data Visualizations.

After presenting my Bachelor's graduation study at a conference, I recall a discussion about the definition of crowd sourcing. The subjects was whether or not unconscious or inexplicit participation in enriching the companies information or resources, was part of the definition. I did include it in my definition. Nowadays, this conflict is solved: this matter is commonly called Big Data.

Digital publishing starts to replace paper publishing. Now real time publishing technology is moving toward a commodity, the competitive power of digital publishing technology reduces. In this new level playing field, publishers can and need to focus on the creation, coordination and delivery of the best content. In order to do so, they need to have insight in the behavior and consumption of readers. To facilitate publishers to discover this knowledge while not overwhelming them, comprehensive visualizations are important. In this study, I develop and test five visualizations for Big Data at imgZine, a company that creates real time social magazines.

I hope you will enjoy reading the text and watching the graphics.

About the author

Bernard (B.J.) van der Wees holds, after successfully defending this thesis, a Master's degree in Business Information Technology from the University of Twente, the Netherlands. Previously he successfully finished the corresponding Bachelor's program. He has a broad general IT knowledge and is proficient in information architecture and design. He has professional experience in technical computer support/system administration, (IT) process analysis and political analysis and gained international experience from professional and academic pursuits at the Netherlands Embassy and Technical University in Berlin. In the upcoming years, he will focus on visualizations, start-ups and politics.

About imgZine

The technology firm imgZine is founded 2011 by Marijn Deurloo and Bert Kok. It has a platform for creating real time social magazines, inspired by products like Flipboard, Pulse and Zite. The platform supports customers in publishing their own customer branded magazines. These are continuously filled with the latest media content, in contrast to more static issue-based magazines. There are all kinds of customers, from journalistic magazines/newspapers to enterprise magazines for internal use (smart intranet, i.e. disclosing the intranet) and external use (e.g. promotional). The magazines built on Business Intelligence and Analytics technologies like an analytics dashboard for publishers and a recommendation engine for readers. (imgZine, 2013a) imgZine is part of ORTEC Living Data, a Big Data lab with multiple start-ups in data intelligence solutions in many different fields. For example, real time bidding solutions for advertisements and sport analytics. (ORTEC, 2013)

Prijsdifferentiatie: One size does not fit all

8 november 2013

Kishan Ramkisoensing
Functie
Industry Analyst

Blog van Kishan Ramkisoensing

Als detacheerder wordt u dagelijks geconfronteerd met de harde werkelijkheid: prijsconcurrentie. Uw innemer of de afdeling inkoop kijkt zeer kritisch naar het tarief en zij eisen in sommige gevallen een daling. En om competitief te blijven, wilt u uiteraard een concurrerend tarief kunnen neerleggen bij uw (potentiële) relaties. Prijsdifferentiatie kan naar onze mening een oplossing bieden.

Differentiëren

In veel gevallen wordt de 'traditionele' tariefsberekening gebruikt om tot een uurtarief te komen. Hier wordt het brutoloon van de werknemer als basis genomen en verhoogd met een factor x om de kosten te dekken en een winstmarge over te houden. Deze manier van berekening zorgt voor een uniforme aanpak voor al uw relaties en er wordt dus in mindere mate rekening gehouden met de actuele en toekomstige markt- en sectortrends, vraag- en aanbodontwikkelingen op het gebied van personeel, concurrentiepositie et cetera. Dit betekent dat u zich soms in uw eigen vingers kunt snijden door niet te differentiëren. In de sectoren waar er nog schaarste heerst (bijvoorbeeld bij sommige IT-specialismen), past een hogere prijs, terwijl in zeer competitieve sectoren (zoals algemene professionals) een lagere prijs wellicht passender is. Het is dus van belang om een goede balans tussen de verschillende eindmarkten te vinden en hier uw prijs actief op aan te passen.

Basis op orde

Uiteraard is prijsdifferentiatie niet zonder risico. Uw backofficeprocessen dienen efficiënt en effectief ingericht zijn. In een oogopslag dient bijvoorbeeld duidelijk te zijn welke werknemers tegen een hogere

Example of article view

Acknowledgements

It is perhaps common knowledge that writing a thesis is not the easiest thing to do. I am greatly thankful to anyone who supported me throughout this process, from constructive and reflective discussions to those who reviewed my thesis. And not in the last place, those who helped me to stay happy throughout this time.

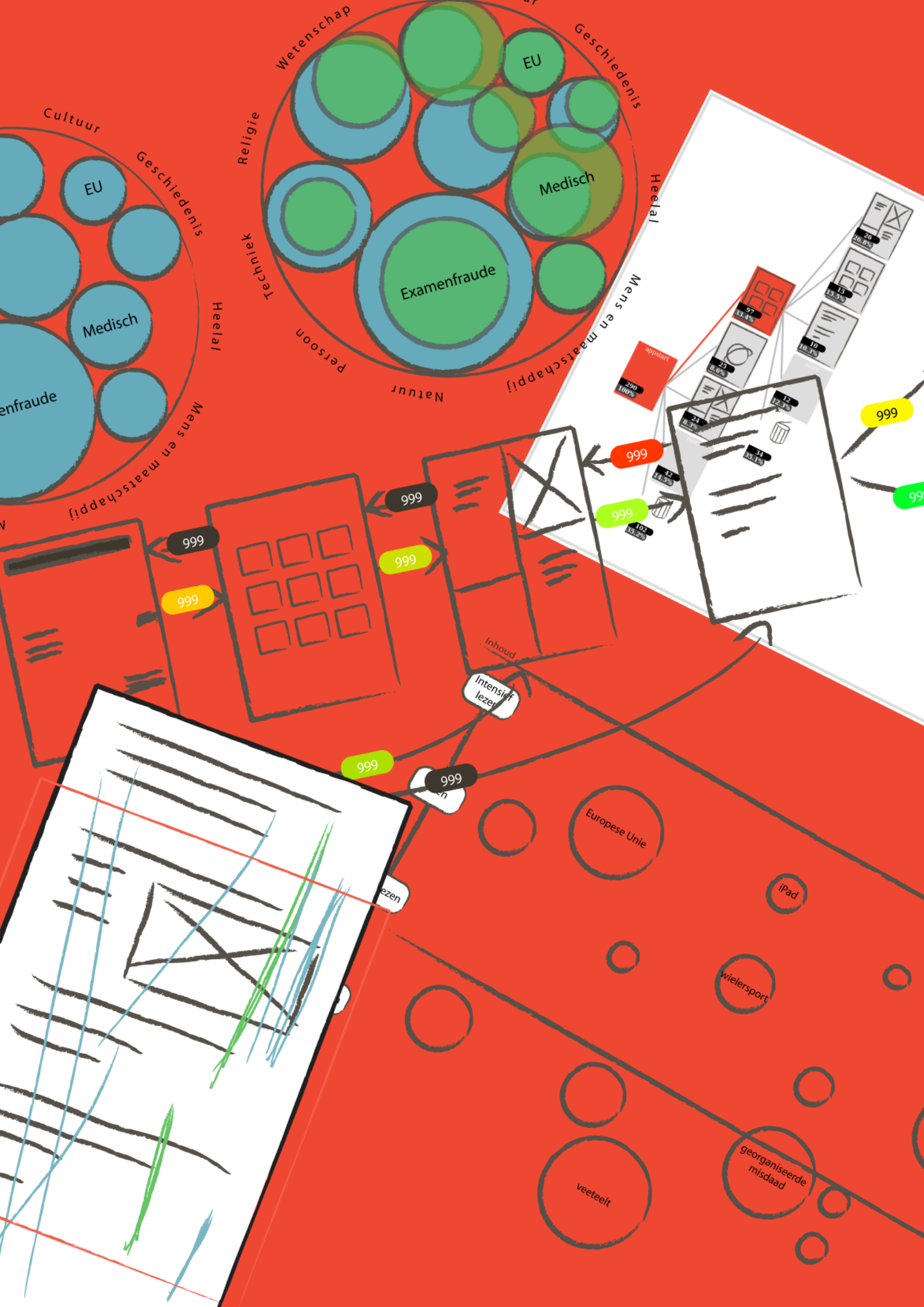
Special thanks go out to my supervisors, Hans Moonen, Maurice van Keulen and Marijn Deurloo for their ongoing support, proof-reading and challenging, yet productive feedback.

Working on this project was absolutely impossible without the support of two brilliant minds, Pedro Silva and Šarūnas Simaitis. Besides laying a technical foundation for my work, some of their mathematics and programming work is presented in this thesis. It was a pleasure to join the 'secret data team'.

Also, I would like to thank all my colleagues for their helpfulness and enthusiasm and support. I would like to give some special attention to Onno Makor for feedback during the visualization process, Thom van Hooijdonk for his feedback on design and products and Gertjan Smits for his feedback on technical aspects and the inner workings of the actual magazines.

Furthermore, I would like to thank my family for their support and positive feedback and social- and user-oriented questions, that helped me greatly to improve my story.

Last, but not least, I would like to thank all those who have proof-read my thesis. You did a good job.


Cultuur
Geschiedenis
EU
Medisch
Heelal
Mens en maatschappij
Wetenschap
Geschiedenis
Religie
Techniek
Heelal
EU
Medisch
Examenfraude
Mens en maatschappij
Natuur
Persoon
999
Inhoud
Intensief lezen
Europese Unie
iPad
wielersport
georganiseerde misdaad
veeteelt

Management Summary

Where

The technology firm imgZine has a platform for creating real time social magazines. The platform is inspired by real time publishing technology like Flipboard, Pulse and Zite have, as well as traditional style magazines on paper. Combing these, imgZine delivers the technology for beautiful digital magazines, that build on Business Intelligence and Analytics technologies like an analytics dashboard for publishers and a recommendation engine for readers. The magazines are continuously filled with the latest media content (real time publishing), in contrast to more static issue-based magazines. Customers are both traditional publishers and enterprise publishers. As for the customers, the magazines vary from respectively journalistic magazines/newspapers to enterprise magazines for internal (e.g. disclosing the intranet) and external use (e.g. promotional).

Why

In recent years, publishers have focused on technology in order to gain competitive advantages. Now that more and more digital publishing solutions like those from imgZine evolve, digital publishing technology is developing toward a commodity. In this situation, publishers can focus the creation, coordination and delivery of the best content and general experience to their readers. As the technology alone does not make the difference anymore, the platform needs to deliver value to the publishers by enabling them to analyze content and discover knowledge about their readers and the readers' behavior. As the publishing technology evolves, it also becomes more complex, as will the questions of publishers. imgZine thus needs to deliver the best general experience and tools with their platform in order to stay competitive in the future. The question rises how the publishers can best be facilitated while using this analysis.

How

At imgZine a new analytics dashboard for publishers is being developed to disclose the content and reading behavior data analysis technologies developed. It is important to facilitate publishers in knowledge discovery, while not overwhelming them. In this research, better suited ways of visualizing are explored by developing novel and adapted visualizations. In a design science study, five artifacts for Information Visualizations of Big Data are developed and evaluated with publishers. Of these five, three artifacts focus on insight in reading behavior/usage patterns on different levels and two focus on insight in content published and consumed. One is designed into a completely functional prototype.

What

Based on our design and evaluation, three main conclusions are drawn:

First, it is found that the detail of information publishers want to have access to, is highly diverse, but reaches from basic statistics to very detailed reading behavior information. This spans the complete tasks of visualizations of Shneiderman (1996).

Second, each of these levels need their own relevant visualizations to maintain the overview by the publishers. The three reading behavior artifacts (Enriched View Trail, Functional User Flows and Swipe Patterns) can be integrated to achieve this for reading behavior.

Third, visualizations that resemble the end-user product clearly, received well understanding among publishers. It is foreseen that the distinction between Information and Metaphor Visualizations (Lengler & Eppler, 2007) will blurry, as more data abstractions are added. It is expected that more new and for a specific purpose designed visualizations are likely to evolve in the future.

Contents

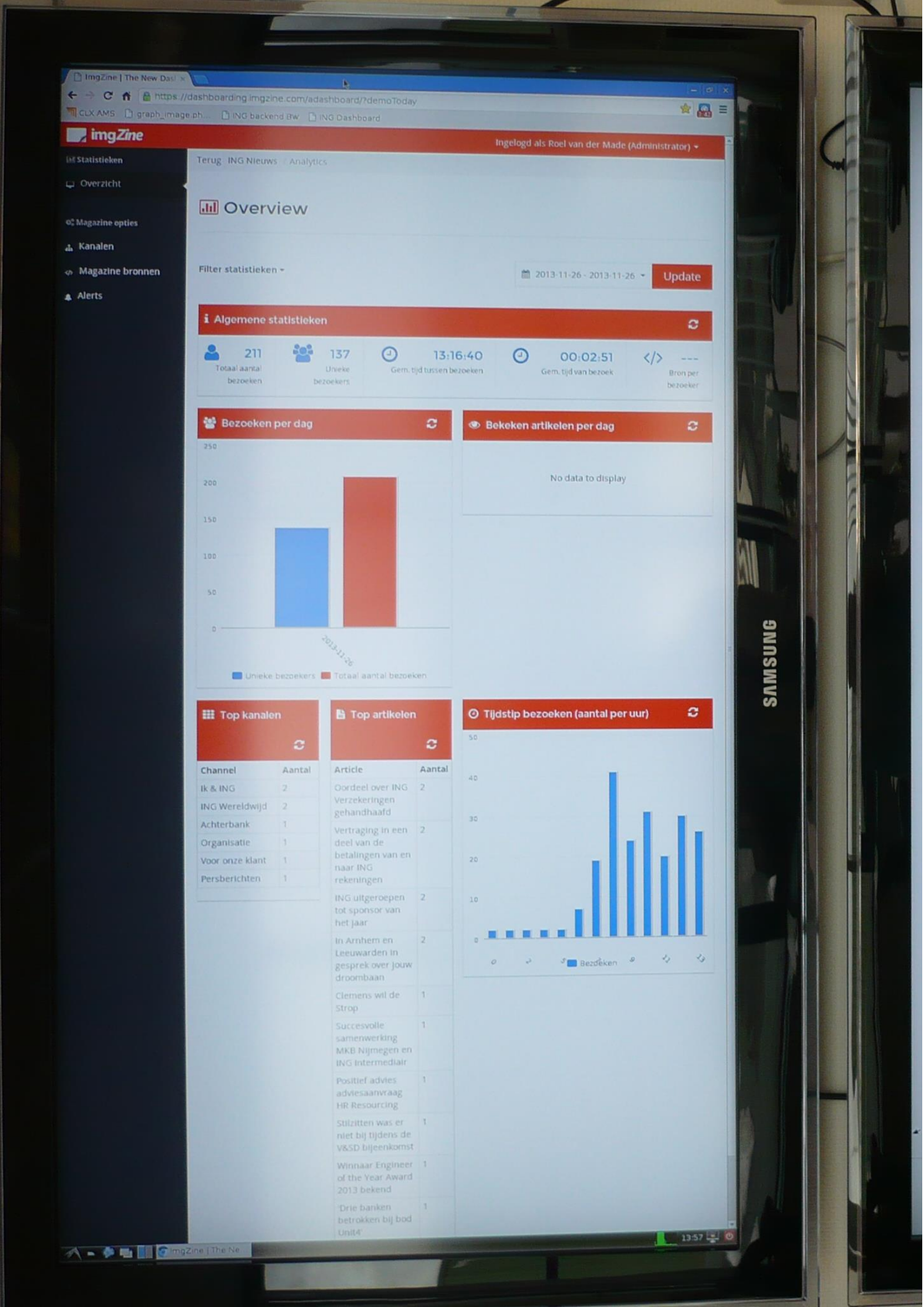
List of Figures and Illustration

List of Tables

List of Definitions and Abbreviations

This is a selection of definitions and abbreviations; the complete list can be found in Appendix A.

Analytics Dashboard	The part of the publishers dashboard where information and statistics about usage can be found. The analytics dashboard and it's users are the problem context of this research.
BI&A	“Business Intelligence & Analytics” (H. Chen, Chiang, & Storey, 2012)
Data Visualization	“visual representations of quantitative data in a schematic form” (Lengler & Eppler, 2007)
Enterprise publishers	Customer group of imgZine, who create magazines for their own employees, or for their customers, most often for information delivery. See also *traditional publishers*.
Information Visualization	“the use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by the user as they proceed working with it” (Lengler & Eppler, 2007)
Publishers	traditional publishers or enterprises who bought their own magazine at imgZine. These are the primary users of the dashboard, configuring their magazine in the configuration dashboard and retrieving information and insight in the analytics dashboard. Two types of publishers are distinguished: *traditional publishers* and *enterprise publishers*.
Real time social magazine	A (digital) app, often multi-platform, which loads articles from sources, as soon as they are released (in case of push), or with a small delay (in case of pull).
Traditional publishers	Customer group of imgZine, who create magazines for external readers. Examples are magazines and newspapers. See also *enterprise publishers*.
UEI	See *User Engagement Index*.
User Engagement Index	A combination of metrics used to determine the actual engagement a user has with a certain article or a set of articles. The basic metric is calculated by correcting the reading speed for the article length. It is used both internally, but also as relative value to the customer for distinguishing the amount of attention. It does not say anything about if the user actually likes the article or not.


ImgZine | The New Das
https://dashboarding.imgzine.com/adashboard/?demoToday
CLX AMS
graph_image.ph...
ING backend BW
ING Dashboard
imgZine
Ingelogd als Roel van der Made (Administrator)
Statistieken
Overzicht
Magazine opties
Kanalen
Magazine bronnen
Alerts
Terug ING Nieuws / Analytics
Overview
Filter statistieken
2013-11-26 - 2013-11-26
Update
Algemene statistieken
211
Totaal aantal bezoeken
137
Unieke bezoekers
13:16:40
Gem. tijd tussen bezoeken
00:02:51
Gem. tijd van bezoek

Bron per bezoeker
Bezoeken per dag
250
200
150
100
50
0
2013-11-26
Unieke bezoekers
Totaal aantal bezoeken
Bekeken artikelen per dag
No data to display
Top kanalen
Channel
Aantal
Ik & ING
2
ING Wereldwijd
2
Achterbank
1
Organisatie
1
Voor onze klant
1
Persberichten
1
Top artikelen
Article
Aantal
Oordeel over ING Verzekeringen gehandhaafd
2
Vertraging in een deel van de betalingen van en naar ING rekeningen
2
ING uitgeroepen tot sponsor van het jaar
2
In Arnhem en Leeuwarden in gesprek over jouw droombaan
2
Clemens wil de Strop
1
Succesvolle samenwerking MKB Nijmegen en ING Intermediair
1
Positief advies adviesaanvraag HR Resourcing
1
Stilzitten was er niet bij tijdens de V&SD bijeenkomst
1
Winnaar Engineer of the Year Award 2013 bekend
1
'Drie banken betrokken bij bod Unit4'
1
Tijdstip bezoeken (aantal per uur)
50
40
30
20
10
0
Bezoeken
SAMSUNG
13:57

1 Introduction

1.1 Background

1.1.1 Where

The technology firm imgZine has a platform for creating real time social magazines. The platform is inspired by real time publishing technology like Flipboard, Pulse and Zite have, as well as traditional style magazines on paper. Combing these, imgZine delivers the technology for beautiful digital magazines, that build on Business Intelligence and Analytics technologies like an analytics dashboard for publishers and a recommendation engine for readers. The platform supports customers, both traditional publishers and enterprises acting as publishers, in publishing their own customer branded magazines. These are continuously filled with the latest media content (real time publishing), in contrast to more static issue-based magazines. Customers and their magazines come in many flavors, from journalistic magazines/newspapers to enterprise magazines for employees (smart intranet) and external use (e.g. promotional). The articles are loaded from different sources, including content management systems, web pages and RSS-feeds. (imgZine, 2013a)

The research environment is described in more detail in section 1.5.

1.1.2 Why

Now that ever more digital publishing technologies and initiatives are available, the digital publishing technology is moving toward commodity. This creates the opportunity for the publishers to focus on their natural habitat again: the content and the optimized general experience.

That said, during the execution of this research, imgZine has rapidly developed and is constantly changing. One of the major changes is the shift in customer focus from traditional publishers toward enterprises who publish (enterprise publishers). Although some initiatives are successful (Pfauth, 2013), the capital available for investment in the publishing industry decreased and innovation is slowing down or sometimes even stands still (Adformatie, 2013; De Morgen, 2013; Reijerman, 2013; Sanoma, 2013; WPG Uitgevers, 2013). However, another market was found in the field of enterprises. The enterprises publish either for an internal or external audience. When publishing for their own employees, often in the form of a new smarter intranet for disclosing their internal data sources in a better way, they try to increase employee engagement. On the other hand, enterprises use imgZine's platform for disclosing knowledge toward the outside world, i.e. to attract new customers or deliver information to current customers.

Now that these enterprises start to use the products for information disclosure, they start to realize more and more that they also start to be publishers. However, to facilitate the enterprises when publishing, knowledge from the publishing field is crucial to the business of imgZine. A lot of activities of the stakeholders at enterprises and publishers are the same, but an increase in the eagerness and willingness of knowledge discovery and in-depth analytics arises.

When using magazines as smart intranet (combining collaboration, document management, social and decision making features) the role of magazines changes. Publishing is used to improve the employee engagement and to deliver information in a better, more interesting and more inviting way to the employees. At one of the customers, the moments of the day readers visited the (smart) intranet: the peak shifted from mid-day visits toward breakfast hours, pointing at an extension of working time and a greater employee involvement.

Some additional issues arise when building magazines for enterprises. These include Bring Your Own Device (BYOD) and the directions of communication. With BYOD, the devices may differ greatly and thus

standardization and information access, protection and security become issues. With a traditional intranet, information can only be pulled from the network, whereas with digital magazines, the flow of information is in many ways both pull and push: people can post using their own social media accounts, get notifications about new, relevant, content and can read new articles whenever they want to.

In order to improve employee engagement for enterprises, or to improve and serve the readers in the best possible way for publishers and enterprises, a good magazine is important. This does include high quality content, the right content for the right reader and a good overall experience. To achieve this, intelligent software is needed in order to select the right articles for the right reader (i.e. automatic recommendations). And, in turn to control these recommendations, better insight for the publisher is needed, which can also be achieved by intelligent and easy to use software.

Using this research, we want to create insight in the content published, the readers and the reader's reading behavior. Questions include: what type of media content (including all articles, graphics, videos, etc.) works better for which group of customers and how can the recommendations be improved? To facilitate this, a new analytics dashboard for publishers is currently being developed, for which we investigate how novel and adapted visualization can contribute to knowledge extraction in the fields of content, readers and reading behavior.

The data that is represented is typically not one or two-dimensional, but contains many dimensions. Also, the aggregation of different types of data gains importance, in order to deliver a coherent overview. Representing the data in traditional data visualizations, like a bar chart, could possibly hide patterns and thus skip important information. Many examples of this situation exist; one of them is the Analytics dashboard by Google, which is considered overwhelming by some publishers, as we discovered from initial customer interviews.

In order to reduce the change of designing an overwhelming dashboard, which does not contribute to knowledge transfer and discovery anymore, some form of data reduction should be applied. Do not show everything, but only specifically selected data. Data can be plotted into a certain format in order to get a better overview of what is done, for example locations instead of ip-addresses. We propose to achieve this data reduction by using information visualizations that create an overview of content, but also allow to dive deep into the content. These should be seen in addition to existing data visualizations for the lowest levels of data.

The research motivation, objectives and questions are described in more detail in section 1.1 and 1.3.

1.1.3 How

We execute a design science study and validate our artifacts with expert sessions and customer interviews. In the design science study based on Peffers et al (2007) with a Marchand and Peppard (2013) approach, five artifacts for disclosing usage behavior are developed. Of these, three artifacts focus on insight in usage patterns and two focus on insight into content. An overview is shown in Figure 1.1.

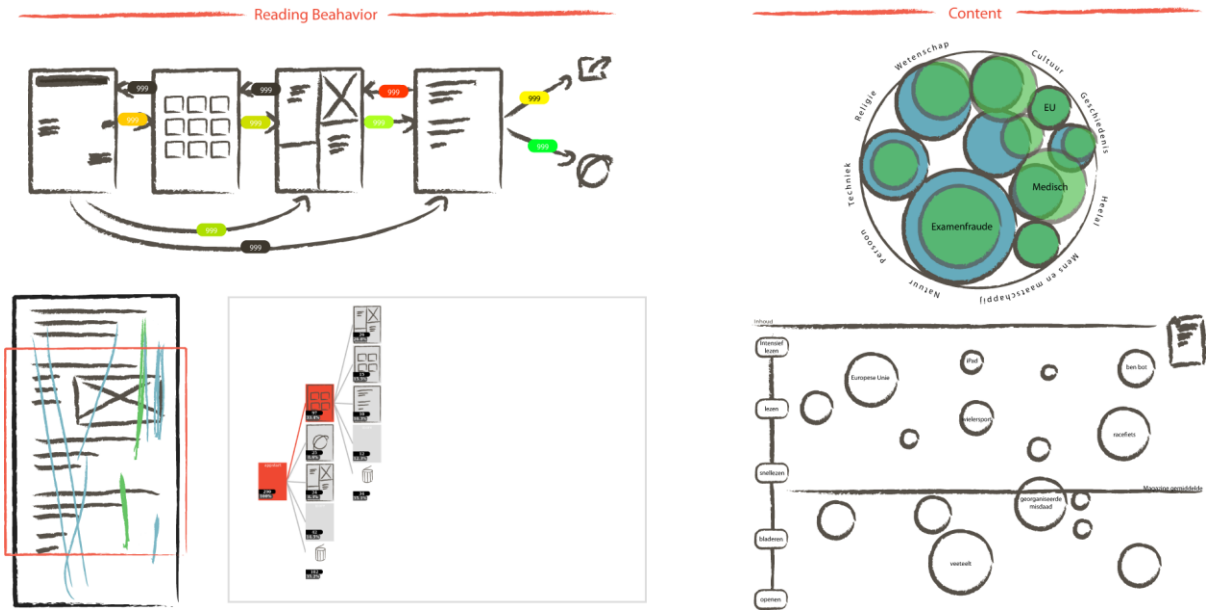


Figure 1.1: Overview of the artifacts

The three artifacts on usage patterns each focus on different levels: Enriched View Trail for a complete overview, the functional User Flows for a more detailed view and the Swipe Patterns for a lowest-entity view. The two artifacts focused on content, Topic Bubbles and Bubble Up, both try to create an overview. The artifacts oriented on displaying readers were not mature enough, to make it to the conceptual phase.

The literature is discussed in chapter 2, the methods in chapter 3 and description of the artifact and the design of the artifacts in chapter 4.

1.1.4 What

The artifacts are evaluated internally and with publishers, in order to validate what they mean to customers. Finally, main conclusions are drawn, which can be found in section 1.1. Next, the limitations and future directions for research and development are discussed.

The evaluations of the artifacts are described in chapter 5, after which the conclusions and discussions are found in chapter 6.

1.2 Research Motivation

1.2.1 What is the Problem?

In order to facilitate the publishers in their work to create the best magazines for their readers, it is important to facilitate knowledge transfer from imgZine's platform toward the publishers. imgZine has a large amount of data about content, readers and reading behavior available and it is growing fast and steadily. Some information is entered by the readers themselves, other is aggregated from usage information presented by the readers' behavior. This data set has a great potential for creating better customer understanding, better insight in published content and thus better magazines and even new business opportunities.

However, the data that comes from the magazines is often multi-dimensional. Especially reading behavior is not only defined by space and time, but also by a lot of other factors. Combining the data to answer even more abstract questions, leads to even greater dimensions. For example, when using complex combinations like association rules ('people who have read article X, have also read article Y') it is important that clear communication is conducted.

Thus, we formulate the following **problem statement**:

> *Publishers need easy and intuitive access to digital publishing technology, so they can deliver the best content and improve reader engagement in order to stay competitive*

1.2.2 Why is it a Problem?

Now that the technology advances, the magazines created get more complex, diverse and personalized. Publishers, both traditional and enterprise, do not always realize they need the knowledge that can be discovered using the platform in order to stay in control. They need the insights from the usage of the magazine in order to control their content and exploit future business opportunities.

Over the last years, many traditional publishing firms have struggled to survive, while trying to compete on technologies. In order to survive, publishers have to adapt to the new situation. When publishing externally, they have to use the knowledge discovered in order to stay competitive. When publishing for employees, it is important for achieving higher employee engagement; something that likely is even more important now that the labor market is increasingly flexible (TNO, 2013).

Furthermore, many traditional publishers are not educated or trained in creating technological solutions, let alone in data analysis or mining. By freeing them from technology and data problems, they can concentrate on their natural habitat again: the overall creation, coordination and delivery of the best content to the right reader.

Thus we formulate the following **problem motivation**:

1. *Digital publishing technology enables publishers to better understand their content, readers and readers' behavior.*
2. *Publishers are not skilled or trained to use the digital publishing tools they need.*

1.2.3 What Causes the Problem?

The data presented is multi-dimensional and thus harder to represent in textual or traditional data or information visualizations. The human mind can only deal with a limited amount of information, also limited by the bandwidth of the eye (Thomas & Cook, 2005). In order to prevent information overload, every type of

data that has more than three dimensions, needs to have a form of data reduction or visualization in order to be interpretable and usable by humans.

Also, the tablets as we now know them have only recently found their way to the market. For example, the first iPad was released in April 2010. As the medium tablet is rather new, only limited research has been executed in analyzing this field. The same holds to some extend for smartphones.

Simultaneously, the targeted users of the system are publishers, who generally have a limited knowledge about the technology and the data that is used. However, they are generally well skilled in identifying their target audience by analyzing information (knowledge discovery).

An important part in the process of communicating and controlling information, is visualization. (Hibbard, 2004) Both on the field this type of multi-dimensional data, the particular application domain (mobile devices and especially tablets) and the research context (the digital publishing industry), not much research toward Information Visualization has been executed.

Thus we formulate the following **problem cause**:

> *It is difficult to create intuitive access to multi-dimensional data with traditional data/information visualizations.*

1.2.4 What are the Alternatives?

Generally, five categories of alternative solutions are identified:

Using **education**: all people using the dashboard have to be skilled in order to be able use the numerical or graphical representations, in order to retrieve the information needed for their knowledge questions. This can be done using documentation, or specifically for this goal designed trainings. This way people would be able to understand complex data sets and graphs. However, both variants are extremely time intensive and thus expensive. Also, every time a user at a customer leaves the company, retraining for the new people would be needed. Creating an accessible dashboard is a good alternative.

Using **commonly used visualizations**. There are unlimited ways of visualization, of which two types are an alternative to developing own visualizations: first, using traditional Data Visualizations, as are used in imgZine's old analytics dashboard. Data Visualizations are directly representing the data. This does not disclose the advance features available, especially regarding reading behavior. This way, important competitive edges are left untouched. Out of our initial interviews, we found that some publishers do and some do not yet realize this. Second, using traditional Information Visualizations. In Information Visualization, an interpretation is added to amplify the cognition. These might be able to communicate the information, but no specific Information Visualizations for reading behavior are found in the literature review.

Using **software from existing vendors**. Solutions like Google Analytics and Appcelerator Analytics could be used to create an overview of how the magazines are used. However, they do not come with solutions for displaying reading behavior or other features that are magazine-specific. Therefore are not able to represent the reading behavior data, as it is specific to imgZine's magazines. Next to that, it was found that Google Analytics is considered overwhelming by some publishes. It should, however, be noted that companies like Google are also doing research toward better and more specific Information Visualization. (Google Ideas, 2013)

Using **Scientific Visualization**. In Scientific Visualizations, the data determines the visual representation (Thomas & Cook, 2005). Although this type of visualization might help for data mining and developing new functionality based on the data, it is not well suited for the domain. The main reason is that the targeted

customer group is new to the field of data mining. Scientific Visualization is the opposite of Information Visualization, as in Scientific Visualization the visualization is *given*, whereas in Information Visualization it is *designed* (Munzner, 2008).

Using **reporting**: imgZine can use their own educated people in order to create reports and send them to publishers. This is costly, but also not future proof, as it would be hard to allow publishers to stay in control over the process of real time publishing, when the information is only periodically delivered to them.

1.3 Research Objectives and Questions

1.3.1 Research Objective and Questions

In order to validate our construct, we explore the possibilities for disclosing data characteristics and patterns by analyzing the data available at imgZine. We try to make optimal use of the data by finding suitable data visualizations and developing new ones that specifically suit our needs and match the data we want to represent. To do so, we execute a design science research with the Analytics Dashboard for publishers at imgZine as the object of study.

Starting from this point, we formulated our research objective and questions. We have formulated the following **research objective**, according to the format proposed by Wieringa (2013):

> Improve imgZine's analytics dashboard *(context)*
> by advanced visualization of readers, reading behavior and content quality *(artifact)*
> such that it is clear and easy to use *(constraints)*
> in order to give imgZine's customers (both traditional and enterprises publishers) insight in their content, readers and reading behavior. *(goal)*

To help us reach our objectives, we formulated the following **research questions**. Our main research questions will be:

RQ: *How can we give publishers insight in their published content, the readers and their reading behavior in real time digital social magazines?*

From this starting point, we derived a couple of **sub-questions**. Each of the sub questions represents one pillar of this research. SQ1 represents Big Data, SQ2 represents the Visualization and SQ3 represents the Customer Validation.

SQ1: ***What*** *data about content, readers and reading behavior is already available, or can be (easily) acquired, from the imgZine magazine platform and how can we extract this data?*

SQ2: ***How*** *can we visualize this data in an easily understandable and accessible way?*

SQ3: ***Why*** *and to which extent do these visualizations contribute to knowledge discovery by publishers in their magazines?*

1.3.2 Guiding Questions

In order to make it easier to answer our research questions and to create a bridge between the artifact to be designed and the sub research questions, a set of guiding questions is formulated in a small internal brainstorm. These guiding questions are explicitly formulated knowledge questions about the artifact.

We sorted the questions here on the categories (content, readers, reading behavior) to which they are the most important in facilitating information. This list is not complete, neither can we assure it is disjoint. Its goal is to help us evaluate and discuss our artifacts in light of the research questions more thoughtful. In order to

do so, the perceived attractiveness to the researcher is rated on a five-points scale from - - (negative) to ++ (positive).

Table 1.1: The guiding questions, their category and perceived attractiveness

Number	Guiding Question	Category	Attractiveness
GQ1	Is there a difference between magazines by younger and older companies?	Content	--
GQ2	Which sources are better read than others?	Content	+
GQ3	Are some articles better not published at all?	Content	++
GQ4	Do certain sources even degrade reader engagement for a complete magazine?	Content	++
GQ5	Could an optimal magazine automatically be composed out of available sources?	Content	+
GQ6	Could sources be selected based on usage pattern?	Content, Reading Behavior	++
GQ7	Can we determine which articles people want to read, but do not read?	Content, Reading Behavior	++
GQ8	Can we determine the age of readers?	Readers	+/-
GQ9	Which type of readers do we have? / Can we distinguish different groups of readers?	Readers, Reading Behavior	+
GQ10	What are the characteristics of these reader groups? / Can we distinguish characteristics like the educational level, income and gender of readers?	Readers, Reading Behavior	+
GQ11	Can we distinguish readers based solely on their behavior patterns through the magazines?	Readers, Reading Behavior	++
GQ12	At which days and times are the magazines most intensely read?	Reading Behavior	++
GQ13	At which locations are the magazines read the most?	Reading Behavior	+/-
GQ14	Do readers make different decisions when reading in landscape or in portrait orientation?	Reading Behavior	+
GQ15	Why do people drop off directly? And how many of them do so?	Reading Behavior	+/-
GQ16	Why do people download an app and never open it at all?	Reading Behavior	+
GQ17	How often do people open magazines without reading anything?	Reading Behavior	+
GQ18	Do people read articles in a specific order?	Reading Behavior	++
GQ19	Do people choose for specific articles to read, or do they pick them (relatively) random?	Reading Behavior	++
GQ20	Can we distinguish different reading behavior between different devices/device types?	Reading Behavior	+
GQ21	Can we distinguish working hour patterns? Are these magazine specific?	Reading Behavior	+
GQ22	How and why do people read articles less or more extensive?	Reading Behavior	++

The complete list, including all other rankings made later in this research can be found in Appendix C.

1.4 Scope and Focus

In this research, the focus is on finding/developing visualizations that will help publishers to understand the data and enable them to find insights by using the artifact, while not overwhelming them. Therefore, the focus is on types of data and what can be done with it, over the speed necessary to fluently run these applications in a real life scenario. Such a trajectory is important, but can be challenging, as put in other words by Keim (2002): "The exploration of large data sets is an important but difficult problem. Information visualization techniques may help to solve the problem. Visual data exploration has high potential and many applications, such as fraud detection and data mining, will use information visualization technology for an improved data analysis." Wong and Thomas (2004) identify a comparable trend, the growing importance of dealing with variety and complexity next to the size of data. This is illustrated in Figure 1.2, based upon the pillars of this research. Under Big Data the three characteristics common among definitions (see section) are written down, under Visualization the categories of Lengler and Eppler (2007) and under Customer Validation the ways of collecting data as described by Yin (2008, p. 102).

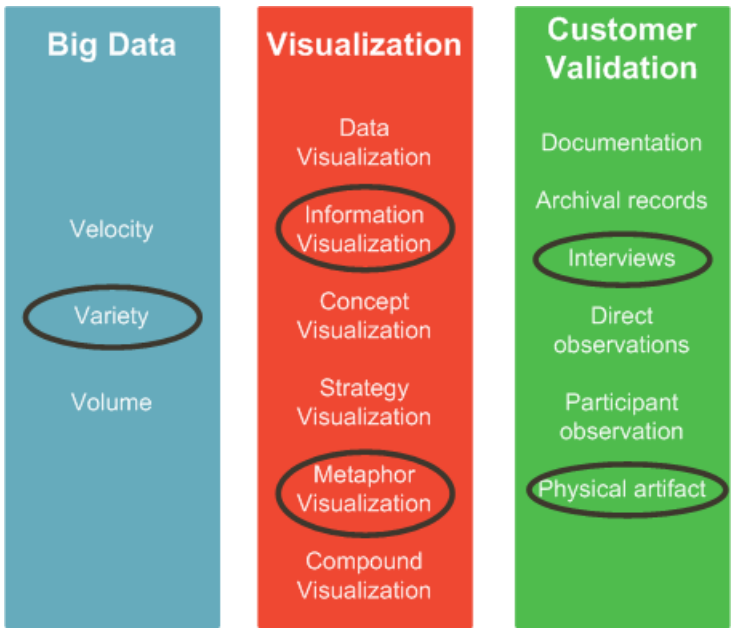


Figure 1.2: The focus and scope of this research

We are focusing on 2D visualizations, due to two reasons: first, the technology to represent 2D visualizations and graphs is better developed and generally more usable. Therefore, a 2D visualization is easier and more likely to be feasible to develop and implement at the moment. Second, it is more likely that publishers are known with 2D visualizations, and therefore easier adapt to new 2D visualizations. Third, according to Munzner (2008), one should be careful with using 3D visualizations; they should only be used when the object it represents, is also a clear 3D surface, for example for representing air flow tests of air planes.

1.5 Research Environment

This research project can be seen as an environment-initiated study (Peffers et al., 2007), as we first defined the demonstration environment, before we identified what the artifacts should look like. The study is executed at imgZine, a young company in the business of creating real time social magazines. It was founded in 2011 and had 12 employees, when this research was initiated, but is growing rapidly: currently there are about 25 employees. The culture at imgZine can be described as very open to initiatives and innovation and not strongly organized and structured. In this environment the structured and planned development of software is unlikely and thus a flexible setup for developing our artifacts and conducting this research is likely to be successful.

Inside of imgZine, the data mining and visualization team consists of four members: a data mining/artificial intelligence expert, an econometrist, an operations/infrastructure architect and the author, for data visualization and concept development. From time to time others join or help the team with their expertise. Such an environment is considered good for designing new technologies and visualizations. (Shapiro, 2010, p. 15)

1.5.1 Real time social magazines

Most digital magazines currently in the market are issue-based. In contrast, the magazines created by imgZine are real time magazines, meaning they contain an ongoing stream of articles and are not issue-based. The magazines are automatically fed with articles fetched from different sources, including a WordPress-installation, website and RSS- or comparable feeds. Also social sources like Twitter, YouTube channels, etc. can be included. As soon as a source has new articles available, it is fetched and pushed to the magazines. This way the curiosity of the reader is aroused, as every time he or she opens the magazine, new articles might be available.

Next to the pushed content, there is an internal selection of articles, based on the interaction with the magazine. Depending on the implementation in the magazine, this is for example a channel containing recommended content, which is different for each user. This also includes articles shared on social networks, saved for later or marked as favorite. The final goal is to create completely, automatic, customized magazines based on both the publisher's decisions and the reader's characteristics and behavior with respect to the magazine.

1.5.2 Magazine Structure

The magazines produced by imgZine for their customers, are sharing a common structure, as can be seen in Figure 1.3. When the magazine is opened the (optional) cover is shown, containing the title, nice graphics and several headlines, containing selected articles and/or channels. The cover is often hidden by the device, in case the application was opened recently on the same device. The index shows the channels grouping sources. A default set of channels is shown, which can be customized by the readers. When a channel is selected, the magazine view shows articles inside of the channel. The different views of the channel are automatically selected and the articles are fitted into different layouts for every page. When the reader clicks on an article in the magazine view, the complete article is shown.

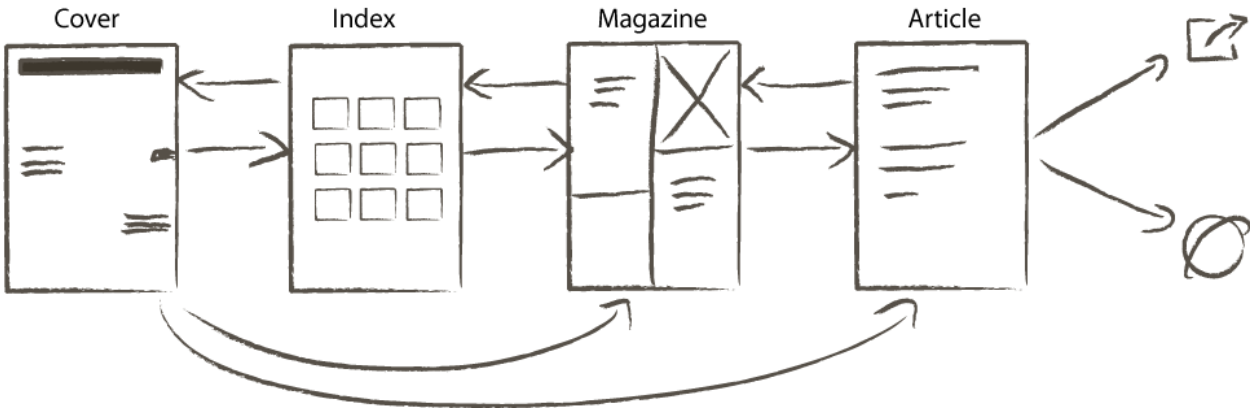


Figure 1.3: Magazine views and structure

An example of how a magazine looks is shown in Figure 1.4. Each magazine is uniquely designed and often has some of its own adjustments, but the core of the magazine is generic. This way the data generated by the magazine about the usage is generalizable to the model presented here.

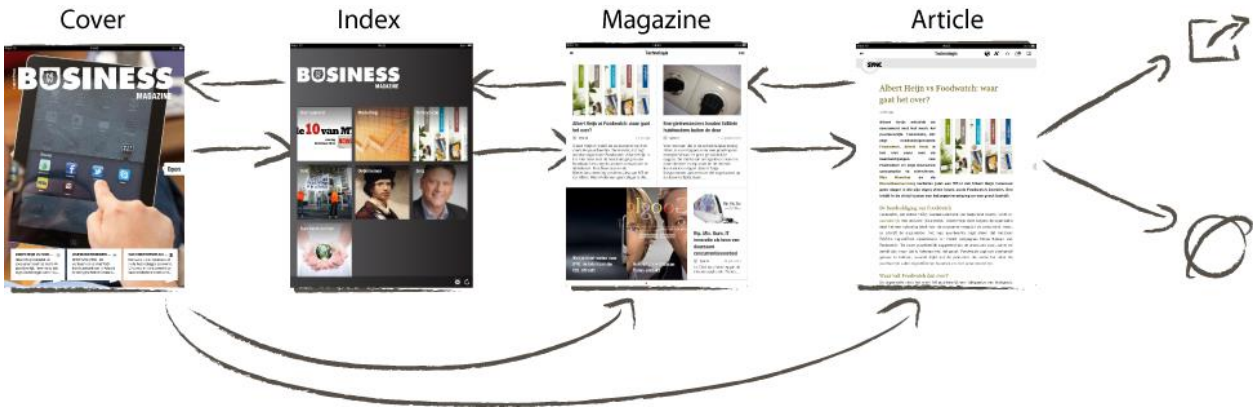


Figure 1.4: Example from the NLinBusiness magazine of how page types look like in production.

1.5.3 Software Structure

The software consists of a single code-base, written in JavaScript on the Titanium Platform (Appcelerator Inc., 2013), which can be exported to multiple platforms, including iOS and Android. On top of the Titanium-base, imgZine works with a common code core, currently called core2, which is a generic empty magazine. Every magazine produced has its own specific configuration and style extensions on top of this core magazine, containing all graphical representations and specific designed pages, like its own distinctive cover page.

1.5.4 Data Team

In order to develop new technologies and intelligence with respect to the data, a data team is composed. This way imgZine tries to explore and exploit future business opportunities. Activities include behavior and text mining, personalization and customer feedback. Examples of developed products are a recommendation engine, the matching of articles with their topics and many more.

1.5.5 The Analytics Dashboard

The magazines imgZine creates are supported by a dashboard for the publishers. This dashboard consists of two functions: the configuration dashboard (for example, defining ‘channels’, which contain articles in the magazine) and the analytics dashboard, which displays insights about the usage of the magazines. Analyzing the configuration dashboard is beyond the scope of our research. The old analytics dashboard shows only traditional usage statistics in data visualizations. These include for example the number of visits, unique visitors, the most read articles and average time between visits. A screenshot of this dashboard can be found in Figure 1.5.

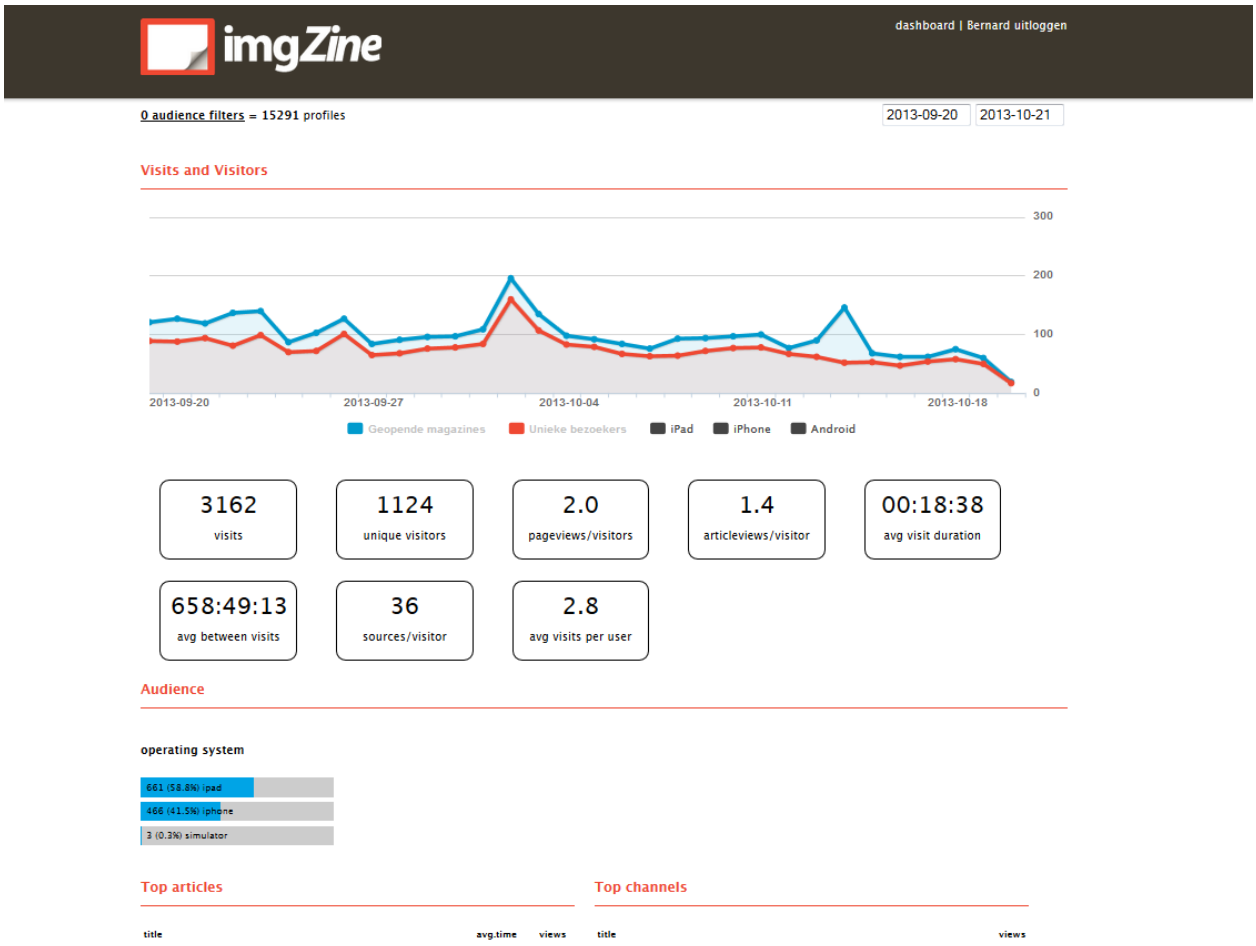


Figure 1.5: The current Analytics Dashboard at imgZine

The newly created dashboard will contain the same basis functionality, but also more advanced features. This dashboard, called the Analytics Dashboard, is the context for which this research proposed novel visualizations. An initial version of the new Analytics Dashboard is shown in Figure 1.6. This is the current state of development and it does not yet contain the designed and investigated visualizations.

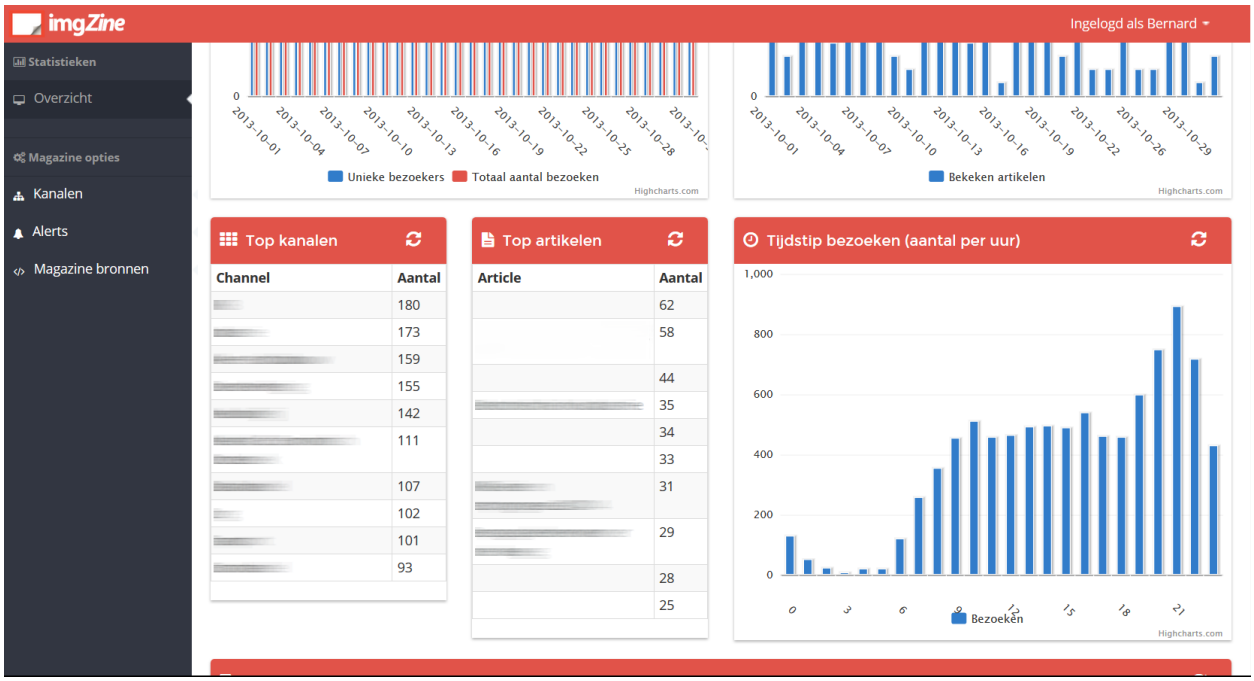


Figure 1.6: Initial version of the new Analytics Dashboard

1.5.6 Customers

The customers of imgZine split in two groups: the traditional publishers and enterprise publishers. Traditionally, imgZine started with creating magazines for traditional publishers, who wanted to bring their paper magazines to a digital platform, or realize a new digital product. As imgZine is a young and dynamic start-up, the focus gradually shifted from traditional to enterprise publishers. Enterprises use magazines for informing employees or customers. This way, they try to disclose information in order to achieve higher employee engagement. Although these enterprise publishers have different goals with the magazine, they execute many of the same tasks as traditional publishers and thus both groups can be seen as publishers.

The enterprise publishers vary in size, but include some very large public and private organizations. Out of the five customers that were interviewed, throughout the research, two are considered enterprise publishers: the Dutch Tax Authority, de Belastingdienst, and an international bank with Dutch roots, the Rabobank. These two enterprise publishers are questioned for the customer validation of the final artifact. The results of these interviews are found respectively in section 5.7.1 and 5.7.2.

1.6 Stakeholders

To create a clear picture of the stakeholders relevant to our research, we identified the relevant stakeholders. Afterwards the most important stakeholders are described in detail. The complete list of stakeholder based on the framework of Clements and Bass (2010) can be found in 0.

(Corporate) Customers	See *Publishers*. All customers are considered part of *traditional publishers* or *enterprise publishers*.
Publishers	The customers of imgZine publish content in their own magazines using the platform designed by imgZine. Two types of Publishers can be identified based on their goals and intended audience. The *traditional publishers* use it as a product and mostly publish for external readers. The *enterprise publishers* use the platform for improving internal information delivery or external relations.
Traditional Publishers	Traditional publishers use the platform for publishing to a broader public and might try to generate revenue directly from the product either by paid reader subscriptions, or by advertisements. Their main goal is thus to generate as much as possible revenue using the product. To do so, they want to retrieve information about how (much) and the product is used by means of clear usage statistics, understand if content is popular and adept content to customer needs.
Enterprise Publishers	Enterprise publishers use the platform for internal information disclosure, for example as extra intranet. Their main goal is to inform their users or employees and improve customer relations. Most often, it generates revenue only indirectly.
imgZine's CEO	The CEO want to improve and simplify insight in data by better data extraction and aggregation technologies and improved visualizations, to give content providers insight in what content is popular and what is content is missing. He also wants to create opportunities for new products to be cross-sold.
imgZine's COO	The COO want to keep the software architecture workable, and prevent threating daily operations. He also wants to improve the recommendation and information statistics in order to gain and retain a general competitive advantage over the competitors.
imgZine's Project Manager	The Project Manager(s) want to reveal missed changes by customers and enable future cross- and up-selling of products.
imgZine's Developers	The developers want to design challenging technologies, while not interrupting daily operations.

1.7 Thesis Outline

This thesis is reported according the reporting format proposed by Gregor and Hevner (2013). The outline is graphically presented in Figure 1.7, along with how the three research pillars are integrated.

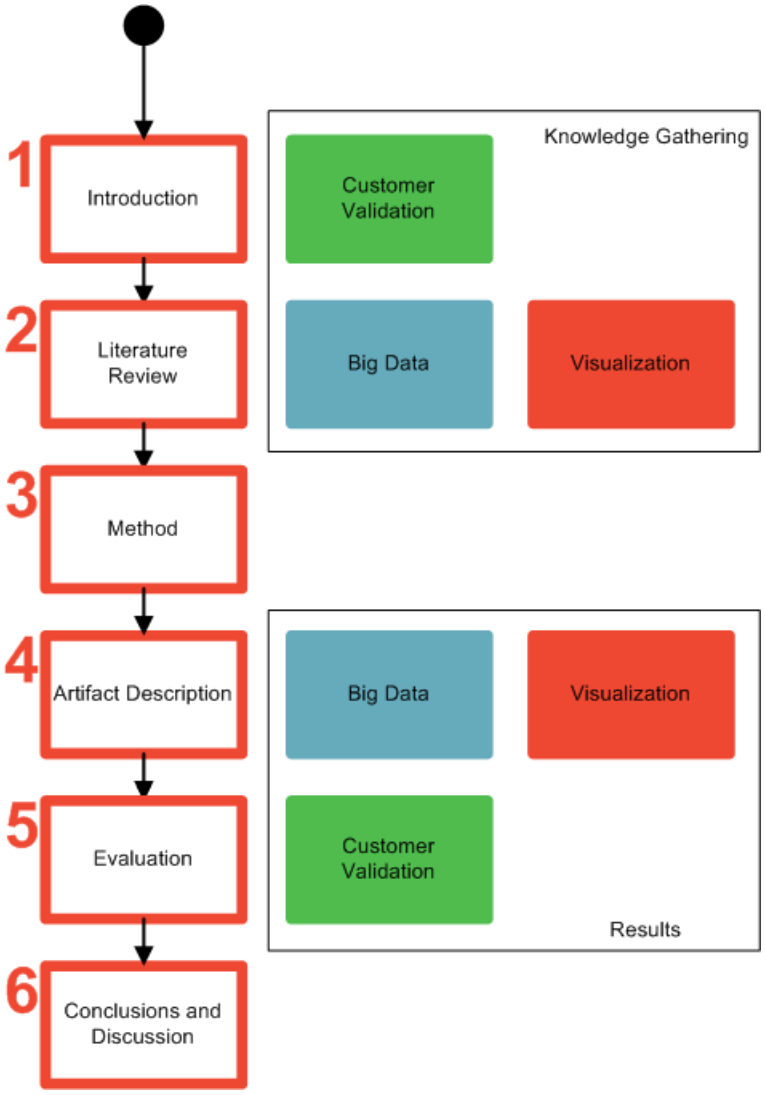


Figure 1.7: Thesis Outline

In this chapter, the research is introduced, the research environment and motivate the reasoning behind this research. The research problem is identified, and the research objectives are formulated, as are the research questions to reach the goals. Last, the relevant stakeholders are identified and research environment is described.

In chapter 2, a literature review is carried out for positioning the research and to gather knowledge about the state of the field as input to our design phase.

In chapter 3, the used methodology, research design and approach are discussed. This research is designed according to the methodology of Peffers et al. (2007). The ways of data collection and the selection of the artifacts are also discussed here.

In chapter 4, the artifacts are described, both from a data and a visualization perspective. The design search is also presented in this chapter.

In chapter 5, the validation of the artifacts is presented, by demonstrating them to the publishers. The results of the expert sessions are also discussed in this chapter.

In chapter 6, the findings are discussed, and interpreted. Sub and main conclusions are drawn, and directions for further research and development are discussed.


Programming Collective Intelligence
O'REILLY
Writing Apache Modules with Perl and C
O'REILLY
Web Client Programming with Perl
O'REILLY
Cookbook
O'REILLY
The Internet Case Study Book
JavaScript
Optimizing Linux Performance
System Performance Tuning
Hadoop
THE LONG TAIL
THE INNOVATOR'S DILEMMA
SEO
PREDICTIVE ANALYTICS
Sander Tamaela e.a.
WILEY
Snapping 3.0
Intranet technologie

2 Literature Review

In this chapter we will look at the current state of the relevant research fields. First we explain our steps and give some general overview of the literature found. Then, we position our research in relation to the research fields found in literature and after that the results for the most important research fields will be explained in more depth. We are specifically interested in what is currently the state of the fields of visualizations, Big Data and reading behavior.

A structured literature review using Google Scholar was executed, related works were examined and peer discussions were held. Using Scholar, we found more literature than reasonably needed for our research, thus we see no need for an even broader search. Information not written in English, Dutch or German is discarded, as are files that are unreachable, or corrupted. For information about the social context, we focus on the Netherlands, as this is currently, the primary market of imgZine – although the focus on the USA is increasing.

Using this method, we found 274 articles of which we read the abstract. Of these articles, 75 are selected and cited in this thesis, of which 31 articles are extensively read. These numbers include articles and information from non-peer reviewed sources. An graphical overview is presented in Figure 2.1.

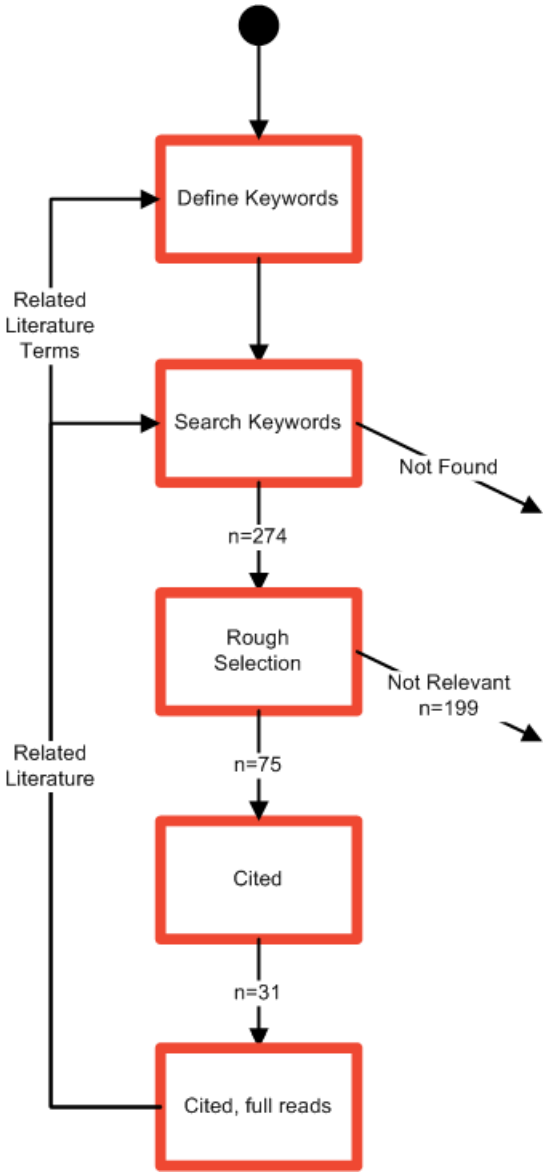


Figure 2.1: Overview of the Literature Review

2.1 Positioning

For a general overview, three research fields have been selected, namely Big Data extraction, Big Data Visualization and Reading Behavior. Two out of three correspond with our pillars; the pillar Customer Validation is represented in the Research Environment. This is illustrated by Figure 3.

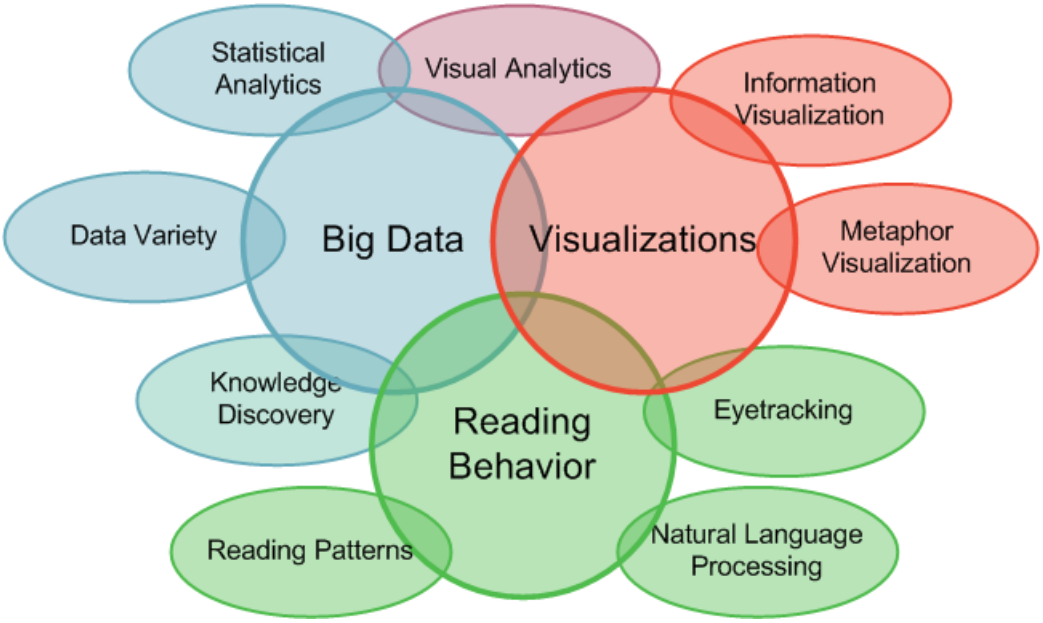


Figure 2.2: Positioning of the research, in respect to relevant research fields

Figure 2.2 presents a Venn diagram with an overview of the fields that are investigated and the relations among these fields. As always, this is subjective to discussion, as some field are highly disciplinary, like visual analytics and a subjective selection has been made, based on the perceived relevance to this research.

2.2 Big Data

2.2.1 Definition, Origin and Evolution

The term 'Big Data' is basically a container definition of aspects related with handling large amounts of data, sometimes using unconventional techniques. Probably most common definitions of Big Data consists of the distinction of three elements regarding data, all starting with a V: *Volume*, *Velocity* and *Variety* (Laney, 2001; SAS Institute Inc., 2013; Zikopoulos, Eaton, DeRoos, Deutsch, & Lapis, 2012). However, the exact definition of the term Big Data is still unclear (Bloem, Van Doorn, Duivestein, & van Ommeren, 2012; Snijders, Matzat, & Reips, 2012). Of course, all aspects of Big Data are relative to the current cost of computing power, data storage, etc. available. Through the years, new data related terminologies and technologies evolved, all by their own terminologies. Nowadays, most of the questions concerned with dealing with great amounts of data are put under the umbrella of Big Data.

Although the definition is relatively general and the term can be used in many perspectives, it comes down to a need for different treatment, where the data is handled more as an asset and less as a by-product of information retrieval, generation and other data processing methods. Also, due to the increase of computer power and sensors, more data is generated and stored, before the actual goal is defined. (Zikopoulos et al., 2012)

The term Big Data can also be described in what it means to business and society, both in terms of technological and social. In this sense it is the process where data given or generated by humans or sensors is stored, before the goal is exactly defined. Data about all different facets of user characteristics or behavior in order to create new insights and by doing so, improve and personalize products and in the end improve the overall experience that users have with regard to a product. (VPRO, 2013)

2.2.2 Evolution Business Intelligence & Analytics

Although the term Big Data deals with problems that are evolving with larger amounts of data, the product created by it in the end evolves to already known OLAP and ETL technologies. Therefore, we see the greatest difference Big Data brings us is the focus on data and what can be done with it (in terms of communication) instead of the focus on technology and tools. (Chaudhuri, Dayal, & Narasayya, 2011; Marchand & Peppard, 2013)

Through the years, we have seen multiple generations of Business Intelligence & Analytics. (Chen et al., 2012) All generations are known for their data-centric approach and have their foundations in the database management fields, but the amount of data, to which extent the data is structured and what the sources are, is different. The generations are visually represented in Figure 2.3.

In the first generation, the content stored was mostly structured and research and development focused on handling data fast and accurately. Traditional DBMS/database manufactures like Oracle have their foundations in this generation. In the second generation, the focus lays on web-based, unstructured content. The amount of data available greatly increases and originates more and more from external sources. Great search engine organizations like Google and Yahoo! evolve. In the third generation, the focus is on mobile- and sensor-based content, delivered in huge amounts and unstructured.

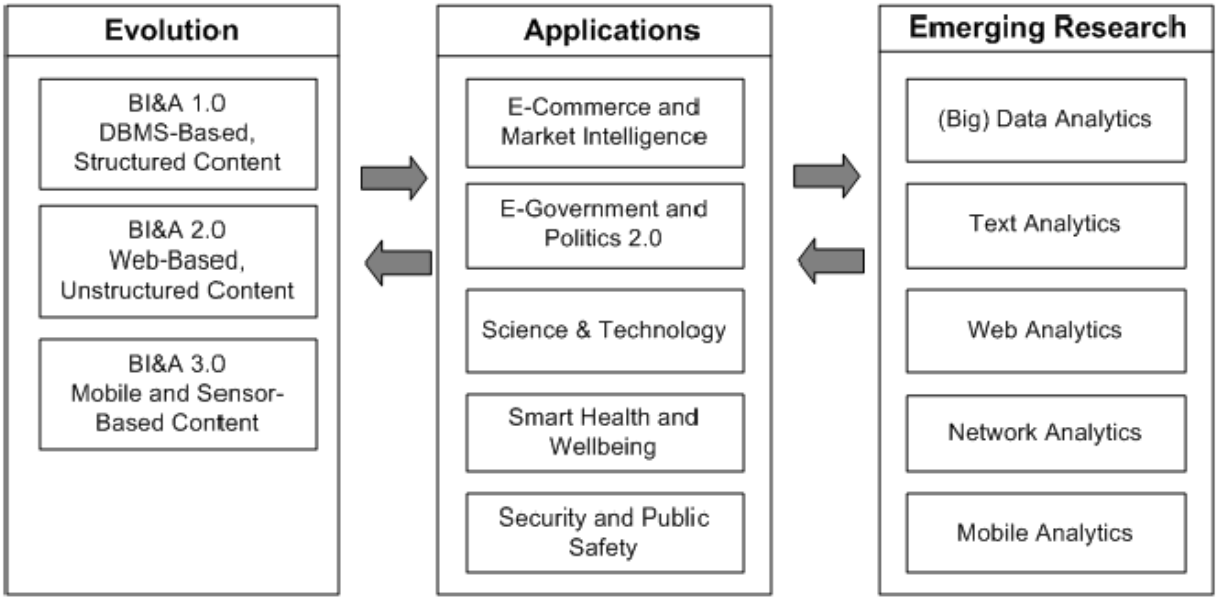


Figure 2.3: Evolution, Applications and Emerging Research in Business Intelligence & Analytics (H. Chen, Chiang, & Storey, 2012)

2.2.3 Challenges

With these large amounts of data, and greater complexity, several challenges emerge. A couple of these are: reducing response time, introducing pretty pictures, dealing with abstraction (Hibbard, 2004), integrating scientific and information visualizations and human-computer interaction (Johnson, 2004). Also in related fields like Visual Analytics, many options are not yet exploited (Zhang et al., 2012). The great lists found in publications and editorials tells us, that many challenges are still ahead of us (Bizer, Boncz, Brodie, & Erling, 2012; Bremer, 2007; Johnson, 2004; Keim, Mansmann, Schneidewind, & Ziegler, 2006; Labrinidis & Jagadish, 2012; Laurila et al., 2012).

2.2.4 Big Data Technology

As storage becomes cheaper, it is easier to store more information, also information of which the goal is not yet clear. Although traditional technologies and methods might work to analyze this data and create information from it, newer technologies can create new opportunities. These include filtering data, improving the accessibility and traceability, but also creating a better sense of the data that is available.

The development of tools related with *Volume* can well be explained in perspective to the BI&A-generations shown in Figure 2.3. The development of relevant tools started with the Google File System and related tools, developed in the early 2000s (Shankland, 2008). One of them is the MapReduce programming model, of which the first and only internally used version is released in 2003. In the MapReduce-paradigm, all data processing is split up in two simple steps. First, there is a Map-procedure for sorting, filtering and comparable actions and second, there is a Reduce-procedure for summary operations, like counting. (Dean & Ghemawat, 2004, 2008) This way processing can be accelerated rigorously. (Agrawal, Das, & El Abbadi, 2011; Y. Chen, Alspaugh, & Katz, 2012)

Since 2007, the Apache Software Foundation develops Hadoop as a central platform for distributed computing and large-scale processing. (Apache Software Foundation, 2012, 2013) Recently, Hadoop gained momentum and it nowadays seems to be the magical keyword for these developments. Other relevant tools include the data warehouse Hive for SQL-style querying (Thusoo et al., 2010) and recently, Facebook released Presto to the public as open source project, as they ran against the limitations of the MapReduce principle. (Novet, 2013)

MapReduce, Hadoop and the other technologies mentioned above deal with scalability issues. As we currently are not yet up to performance issues – running the prototype artifact is possible (but slow) without these optimizations, this research does not focus on deploying and using these technologies.

The second aspect of Big Data, *Velocity*, deals with questions regarding the speed of solutions and handling great volumes fast. This aspect is beyond of the scope of this research.

The third aspect, *Variety*, deals with the great various kinds of data, for example the different types of data, undefined or unclear formats or, in general, inconsistent usage and processing of the data. The variety of BI&A generations 1.0 and 2.0 is present in this research.

2.3 Visualizations

In order to maximize the utilization of available data and turn it into information, it is important to look at the possible visualizations that will represent this information. In order to create a, possibly new or altered, visualization that best matches our data, we take a look into available data visualizations.

A large number of visualization techniques and methods are available, both in academia (Behrens, 2008b; Heer, Bostock, & Ogievetsky, 2010; Lee, 2012) and in practice (McCandless, 2009, 2013; Wiederkehr, Siegrist, Stucki, Gassner, & Schmid, 2013). A lot of attempts to order visualization types have been made (Behrens, 2008a; Heer et al., 2010; Iliinsky, 2010 p. 5; Lee, 2012; Lengler & Eppler, 2007, 2011; Lima, 2011). One of the most complete is in the form of a periodic table (Lengler & Eppler, 2007) – which is, ironically, a perfect example of the violation of good visualization practices (Iliinsky, 2010). Illiinsky points out that a good visualization stands or falls with the link between the data and the representation of this data. “All the other uses of these formats fail to understand what makes them special: their authentic relationship to and representation of the source data” (Iliinsky, 2010, p. 6). The periodic table is designed to show chemical elements – and defines the links between them. An empty spot means something. When it is there; a certain element is yet to be discovered. Such a practice is not the case in most of the other (mis)uses of the periodic table visualization.

2.3.1 Ontology

Before creating relevant adapted or novel visualizations, it is desired to know what visualizations currently exist, according to the literature. The number of ontologies is in comparison with other IT-related research fields relatively small. This may be due to the great diversity of visualizations, making it hard to create a complete and correct ontology. However, some attempts have been made to create a taxonomy. (Lee, 2012; Lengler & Eppler, 2007; Shneiderman, 1996) Most of them are based on the type of data represented. For example, Shneiderman (1996) describes a taxonomy, based on seven tasks and seven data types.

Despite the criticism that has just been uttered, the ontology created by Lengler and Eppler (2007) is one of the most complete. Therefore, this one is used throughout this research, and can be found in Figure 2.4. Six types of visualizations are distinguished:

- **Data Visualization** is the "visual representations of quantitative data in a schematic form";
- **Information Visualization** is "the use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by the user as they proceed working with it";
- **Concept Visualization** means "methods to elaborate (mostly) quantitative concepts, ideas, plans and analyses";
- **Strategy Visualization** is "the systematic use of complementary visual representations in the analysis, development, formulation, communication and implementation of strategies in organizations";
- **Metaphor Visualization** "positions information graphically to organized and structure information. They also convey an insight about the presented information through key characteristics of the metaphor that is employed";
- **Compound Visualization** is "the complementary use of different graphic representation formats in one single schema or frame".

As we want to turn data into useful information and we try to visualize it in a useful way, we focus on the Information Visualization, as is explained in more detail below.

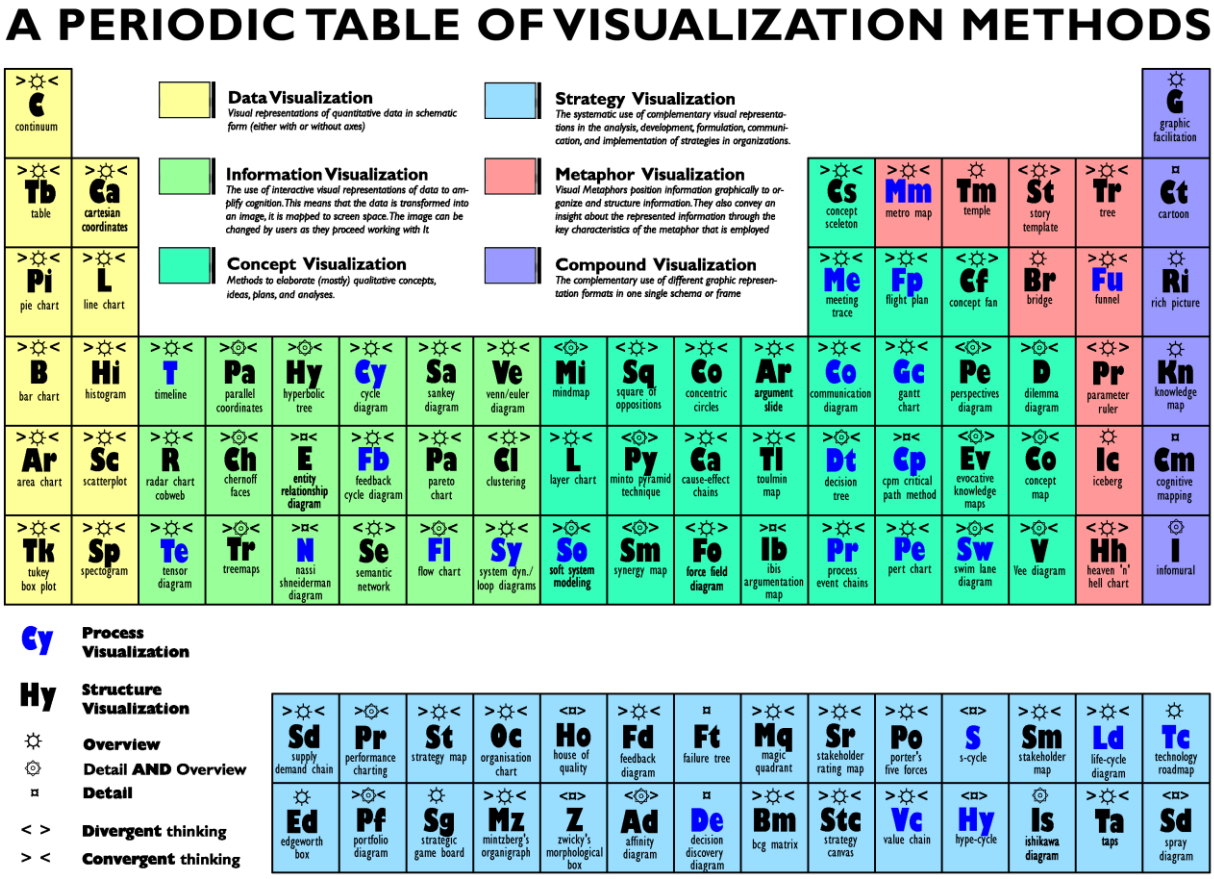


Figure 2.4: Periodic Table of Visualization Methods Source: (Lengler & Eppler, 2007)

2.3.2 Data versus Information Visualization

As previously mentioned, Data Visualizations deal with the "visual representations of quantitative data in a schematic form", whereas Information Visualization deals with "the use of interactive visual representations of data to amplify cognition". This is an important distinction for this research. In Data Visualization, the data is represented directly, without real abstractions and interpretations, whereas in Information Visualization, at least an abstraction is added in the form of a graphical interpretation. Other abstractions, like metrics and data reduction might also be present.

2.3.3 Tasks of Visualization Exploration

There are different views on how data can best be presented for exploration when using visualization techniques. The traditional mantra for diving into data representation is "Overview, zoom, filter, details-on-demand, relate, history and extract" (Shneiderman, 1996). In contrast, Keim et al. (2006) advise a different mantra for data analysis tasks: "Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand." In this research, the mantra of Schneiderman (1996) is used, as the first steps of the mantra proposed by Keim et al. (2006) are resembled by the design phases that are executed in this research: the data is analyzed and important parts are selected. After these steps, both Sneiderman and Keim et al. present comparable tasks.

2.3.4 Design of Visualizations

Many great examples of visualizations are available online. (Visual.ly, 2013) A notable, paper, overview work is Information is Beautiful (McCandless, 2009). In order to design good visualizations, it is important to follow Design Guidelines and Design Patterns. Design Guidelines are found in Beautiful Visualizations (Steele & Iliinsky, 2010). Design patterns help in order to create visualizations that are fast and easily understood. To get insight into design patterns in information visualization, the overview created by Behrens (2008b) is used. McCandless (2010) teaches us that it is important to have relative figures instead of absolute figures, to enable proper analysis and the right overview.

2.3.5 Visual Analytics

A field that is very interesting for our research is Visual Analytics, “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas & Cook, 2005). The field gained attention around 2004, when the National Visualization and Analytics Center (NVAC) was founded by the US Department of Energy. It is a highly multidisciplinary field, as it “combines strengths from information analytics, geospatial analytics, scientific analytics, statistical analytics, knowledge discovery, data management & knowledge representation, presentation, production & dissemination, cognition, perception and interaction.” (Keim et al., 2006)

Most of the research focusses on generic tools (Zhang et al., 2012), usable for many different types of data. This makes it almost inevitable that their ways of representation are not fitted explicitly for the data shown, but are often more general. Zhang et al. (2012) executed a comparative study on the current state of Visual Analytics tools, and present an interesting overview of available tools and their state. Selection of tools is primarily based on market share and further research is executed using structured questionnaires for surveying the vendors and testing with real world data. They found that tools are slow with embracing new visualization technologies: “surprisingly, the number of visualization techniques that are implemented by the surveyed VA systems is rather small compared to the number of techniques that are available from research. ” Simple representations for data with few dimensions, like bar charts and scatter plots, are often implemented, but tools for high-dimensional data, like parallel coordinates visualization, are not implemented often. Finally, they identified five interesting challenges for future directions of Visual Analytics tools, namely: semi-structured and unstructured data, advanced visualization, customizable visualization, real time analysis and predictive analysis.

Bremer (2007) showed that better Visual Analytics tools lead to a reduction of time spent on exploring and understanding trivial data and by doing so allow further search and exploration of in-depth, not evident information and patterns. The research focusses, as most of the Visual Analytics field, on the visualization of raw, scientific data. We focus on tools for visualizing information, thus with interpretation.

2.4 Reading Behavior

Various studies on reading behavior have been executed. The type of screen interface (computer, e-reader, tablet) greatly influences the results. The Nielsen Norman Group (1997, 2006, 2010) for example, executed a lot of research using eye-tracking on especially traditional computer screens. They found, among other things, that people read only little amount of the texts and that what they read has an F-shaped pattern (on top of the page the most is read and down the page the attention shifts to the left).

Interesting research on the influence of design on reading speed (Dyson & Haselgrove, 2001), but important characteristics of the research are not relevant anymore, now that computer screens are not too big to carry around anymore. This brings us toward the tablets, which are expected to have a great influence on magazines, although only recently introduced into the market (Staughton, 2012).


Polish up
your content!
more info?
14:00
dinsdag 26 november
Polish up
your content!
imgZine
real time social magazines
ontgrendel

3 Method

In this chapter the methodology used for this research is described. Firstly, we dive into the structure of analytics projects. Secondly, we discuss different Design Science methodologies and pick our preferred methodology and a reporting format. And last, we look more in-depth at our research model.

3.1 Research Methodology

For achieving our research objectives and answering our research questions, we conduct a design science study. We see a great coherence between the needs of a modern Big Data/Analytics project, as described just before and the process and methodology as facilitated by design science methodologies. A great amount of design science research designs and methods exists in literature, many of which seem to map easily on each other. (Meertens, 2013) Out of the selection examined by Meertens (2013), we decided to pick the methodology from Peffers et al. (2007), based on the clear description of steps and framework and the justification in terms of other design science methodologies mapped on the proposed phases. The framework also facilitates the steps needed for an analytics project, as identified by Marchand and Peppard (2013), very well. They "propose and develop a design science research methodology (DSRM) for the production and presentation of DS research in IS" (Peffers et al., 2007). They do so by identifying common design process elements from earlier design science work and map the design science models by Hevner (2004), Walls et al. (1992), Nunamaker et al. (1996) and others. The outcome framework containing six steps, of which the last five are an iterative process. This process is illustrated in Figure 3.1.

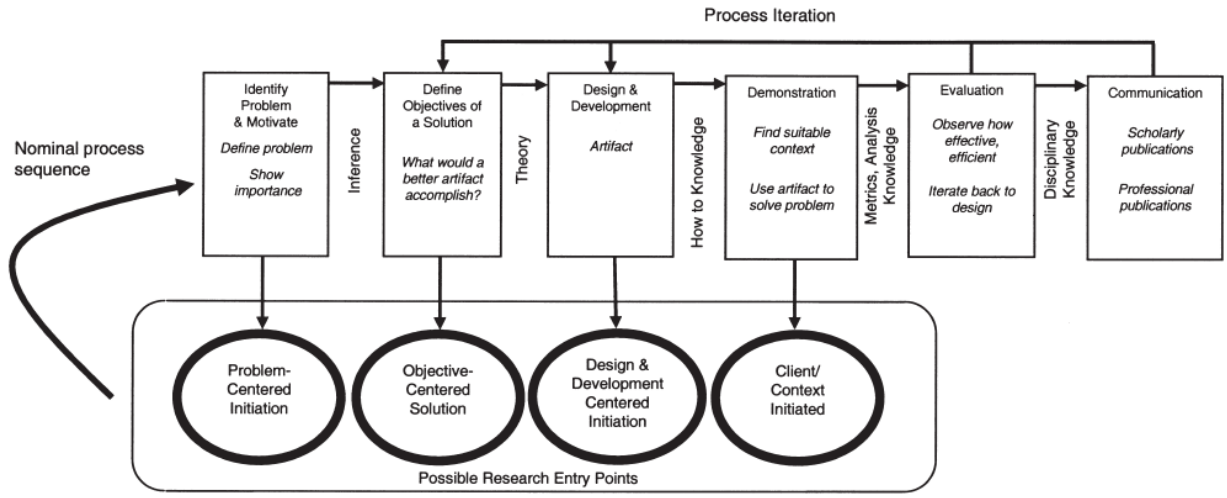


Figure 3.1: The DSRM framework proposed by Peffers et al. (2007)

There are different entry points in the research (Peffers et al., 2007), of which we are in a client/context-initiated project, which means that the environment in which the artefact is demonstrated, has been chosen before the first steps, including the precise identification of the problem, has been executed. According to Peffers et al. (2007) this is a normal situation and does not have to change the description of the project.

In addition to the methodology provided by Peffers et al. (2007), we use Hevner et al.'s (2004) seven guidelines as guiding principles through the design process. These principles are: Design as an Artefact (1), Problem Relevance (2), Design Evaluation (3), Research Contributions (4), Research Rigor (5), Design as a Search Process (6) and Communication of Research (7). These principles are followed, because we consider them helpful in executing solid design science.

Wieringa (2009) points out that Design Science Research is a nested process and a clear distinction between knowledge questions and design problems should be made. He presents eight guidelines, that are in coherence with the guidelines made by Hevner (2004) and earlier. The nested questions are used for guiding and structuring the process, but for clarity reasons only high-level questions are presented in the report.

The findings of this design science study are reported according to the format proposed by Gregor and Hevner (2013), as the chapter scheme is closer to traditional publications and thus allows the reader to understand the structure quickly, while it does not collide with the concepts of design science. The most clear differences between the reporting formats proposed by Peffer et al (2007) and Gregor and Hevner (2013) are that, the design goals for the artifact are less explicitly defined and can now be found as part of the first chapter and that demonstration and evaluation are combined. The chapters conclusion and discussion are combined in this thesis, opposed to the proposed schema.

3.2 Research Model

Based on these frameworks, we created a research design with nested steps, which is presented in Figure D.1: Extensive research design. The main phases are based on Peffers et al. (2007), whereas nested concept is based on Wieringa (2009). Each step represent a knowledge or a design problem, which contains knowledge and/or design problems itself. Knowledge problems implicitly lead to literature review, either on academic literature or on practical knowledge sources. Next to that, two explicit literature review steps are part of the research design.

In five iterations, which are executed partly parallel to each other, we walk through the relevant aspects of our Big Data Exploration and towards a final functional artifact, which is intended as a visualization and technology prototype. For each of the iterations, the research steps from our methodology are followed: Objective, Design and Development, Demonstrate and Evaluate (Peffers et al., 2007). This way, we will touch all relevant aspects in a structured, yet rigid way. This process is illustrated in Figure 3.2 and the phases are described in more hereafter. A more extensive overview can be found in Appendix D.

Five conceptual artifacts are delivered, of which one is further developed as a functional artifact (or working prototype) for further examination. These together are our objects of study, inserted in the research environment at imgZine. Using this research method, we answer our research questions and try to reach our objectives.

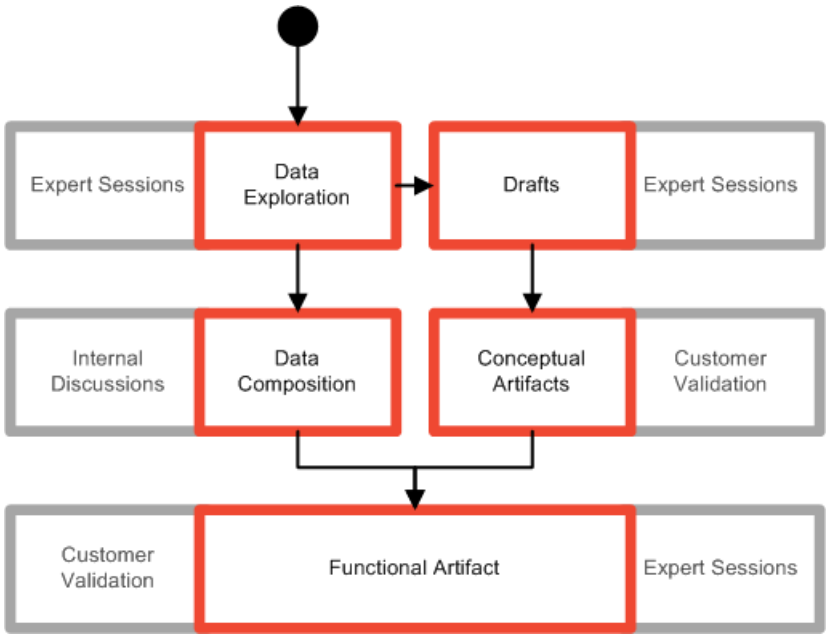


Figure 3.2: The research phases and with their corresponding validation techniques

- **Data Exploration**. Before draft artifacts can be created, an initial exploration of the data available is executed. This phase is continued parallel to the creation of draft artifacts. This phase is validated through expert sessions and internal discussions.

- **Drafts**. In a creative session of about two weeks, a total of 56 possible representations of the data are developed, inspired directly by the data, by visualizations from different sources, by filling gaps in the ontologies found and by combining all of these. This phase is validated through expert sessions.

- **Conceptual Artifacts**. A number of visualization concepts are developed based on the problem identification, knowledge from literature review of currently available visualizations and technologies and using unstructured brainstorming sessions. These concepts are validated in three internal expert sessions, one on technology, one on product and sales and one on usability. In these sessions, the drafts are discussed and based on these results, a limited number of conceptual artifacts is further developed and validated in four customer interviews. This phase is validated through customer validation.

- **Data Composition**. Parallel to developing the conceptual artifacts, the data available and the technology needed to extract this and combine the data into visualizations is researched based on the outcomes of the first cycle will be used as basis for a further artifact, that is for a Technology Prototype implementation. In this data composition also the first steps towards the functional artifacts are executed, by means of touching the relevant technologies and testing the extraction in order to prove that the idea could work. This phase is validated through internal discussions.

- **Functional artifact**. Out of the artifacts developed in the earlier phases, one artifact is selected for further exploration. Based on the inspiration and knowledge from all earlier phases, the artifact is combined and improved for this phase is validation through two customer validations and one expert session.

This research model is chosen, because it is expected to fit in nicely with the culture at imgZine – as described earlier – and match well with the information-focused project structure as described by Marchand and

Peppard (2013). Therefore, it is expected to be supportive to the flexibility needed for this project, as well as that it enables us to learn from earlier steps.

By following this model, development could take place both in a flexible, yet rigid way. By means of the different pre-prototype artifacts, we could assure to be able to learn from the previous steps, as well as that we could focus on all different aspects the artifact to be tested.

3.3 Approach and Project Structure

In current Big Data projects, we have to deal with a great amounts of uncertainty. At the start of a project, it is common that it is not exactly known which data is available and what can be done with the data. In this setting, the traditional approach to developing and implementing IT-systems does not fit the problem. Where traditional IT projects are oriented on the tools and on a corresponding, structured selection process for selecting them, analytics projects should be focused on the data and hypotheses and finding the right representations for using it. (Marchand & Peppard, 2013)

According to the model proposed by Marchand and Peppard (2013), a common traditional IT project is deployed according to the steps "define desired outcomes, redesign work processes, specify technology needs, develop detailed plans to deploy IT, manage organization change and train users and implement plans", whereas Big Data Analytics projects should be deployed using a different project structure, more data-centric. This difference is illustrated by the image in Figure 3.3. (Marchand & Peppard, 2013) As we identify our project as an example of a Big Data/analytics project, we follow the proposed project structure. The project structure they propose, roughly follows the steps hereafter. We map each step to the phases of the research model and explain what it means in this research.

- **Develop theories.** Early in the research and based on the initial customer interview, the goals and characteristics of the customers are identified. Based on this information theories are developed and goals are defined for the research.

- **Build hypotheses.** Based on the research objectives and goals, a list of guiding questions, about the artifacts to be, is composed. By designing the artifacts, as many as possible questions are tried to be answered, without creating information overload.

- **Identify relevant data**. In the Data Exploration and Data Composition, we researched the data available and researched how they could or could not be extracted. Also some expert sessions and internal discussions helped us deliver the right data and find out the relevance of technologies.

- **Conduct experiments**. In the phases Data Composition and the Functional Artifact phase, we experiment with the combination and realization of data and in the different validation sessions, we experiment with the artifacts designed in order to test their ability to help understanding insights, towards our target audience.

- **Refine hypotheses in response to findings**. Based on the findings from our Conceptual Artifacts and from the Data Exploration and Composition phases, the guiding questions and the Functional Artifact is improved based on these findings.

- **Repeat the process.** The research is setup according to the design science methodology DSRM by Peffers et al. (2007). This methodology does enable both freedom for creative design while supporting an iterative process.

An overview of these different phases can be found in Figure 3.2 and a more detailed diagram of how the different methodologies combine, can be found in Appendix D.

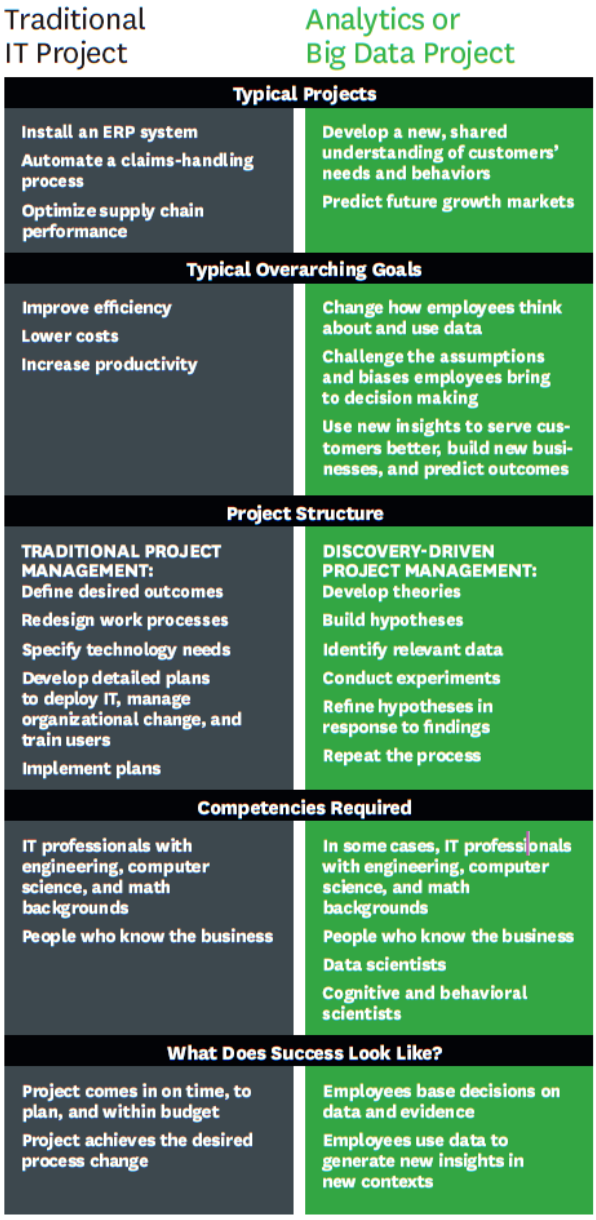


Figure 3.3: Traditional IT project versus analytics project. (Marchand & Peppard, 2013)

3.4 Collecting Data

The data needed for executing this research is collected based on a number of different methods; most of them are qualitative. Statistical analysis is hard to execute, due to the dynamic and fast-changing research environment and the limited number of customers. Combined with the explorative character of this research, data collection by qualitative methods is found to be most suitable. As we try to learn from our previous phases, data is directly collected during or after each research phase.

An overview of how the data is collected is presented in Figure 3.4. Each pillar in this graph corresponds with one of the sub research questions. From left to right: Big Data is related to SQ1, Visualization to SQ2, and Customer Validation to SQ3. An overview of the used data retrieval techniques per relevant steps of Peffers et al. (2007) is also found in Figure 3.4 (the horizontal bars). An extensive overview is found in Appendix D.

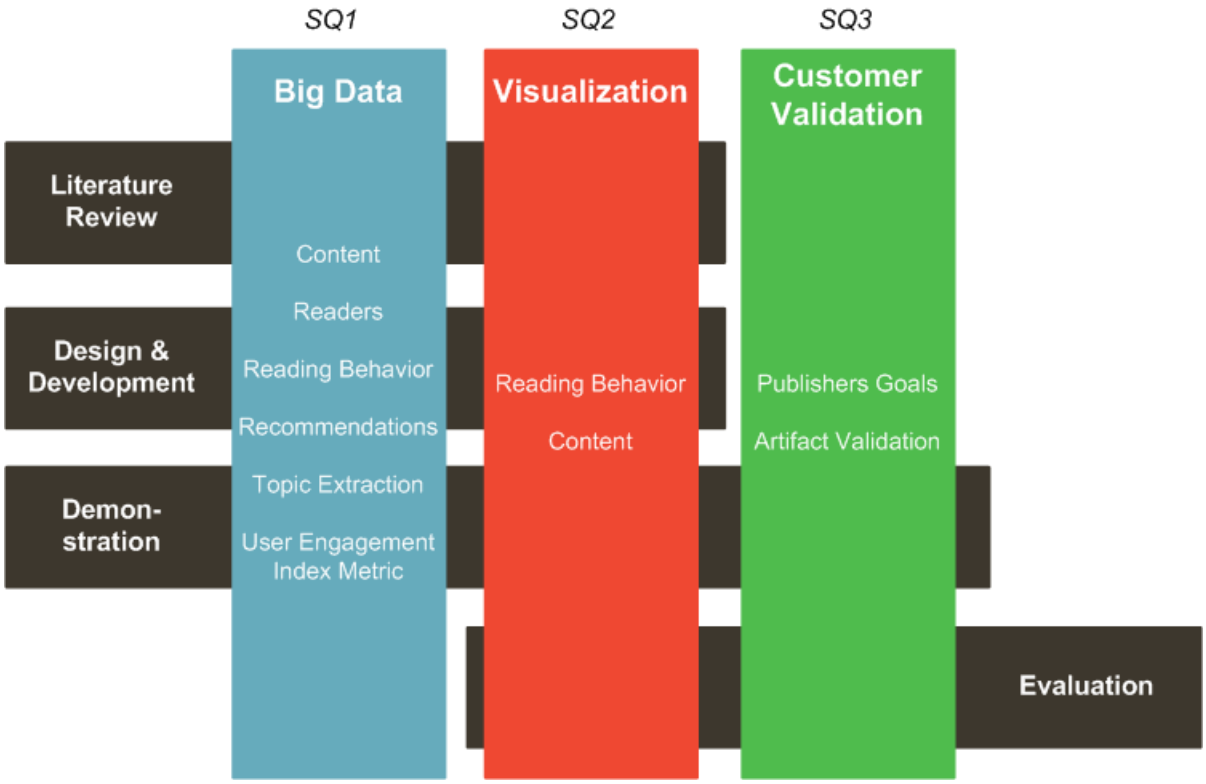


Figure 3.4: Overview of the process of data collection

First, a rough **literature review** is executed in order to get an overview of the current state of research. During the different iterations, more literature is collected as soon as relevant. The methodology is explained in more detail in chapter 2.

Second, in **design & development**, knowledge is gathered by designing and developing drafts, conceptual and functional visualizations. Every step in this process is documented, and all relevant versions of the artifacts are stored, in order to prove reproducibility.

Third, **demonstration** is done by internal Expert Sessions and internal discussions of the (extracted) data and the created visualizations. Next to that, the artifacts are demonstrated to the customers in the evaluation phase during interviews.

Fourth, the **evaluation**: to retrieve information about the customer perception of the product, semi-structured customer interviews are used. These interviews consisted of a general part and part with questions per artifact. In order to reduce bias, all questions are open questions. For a general overview of the procedure to follow, Yin (2008) is consulted. When the customer agrees and technology worked conversations are recorded. This is the case for four out of the total of seven interviews. The main reason for not recording was the technical inability combine recording with executing the working artifact, due to an operating system related problem.

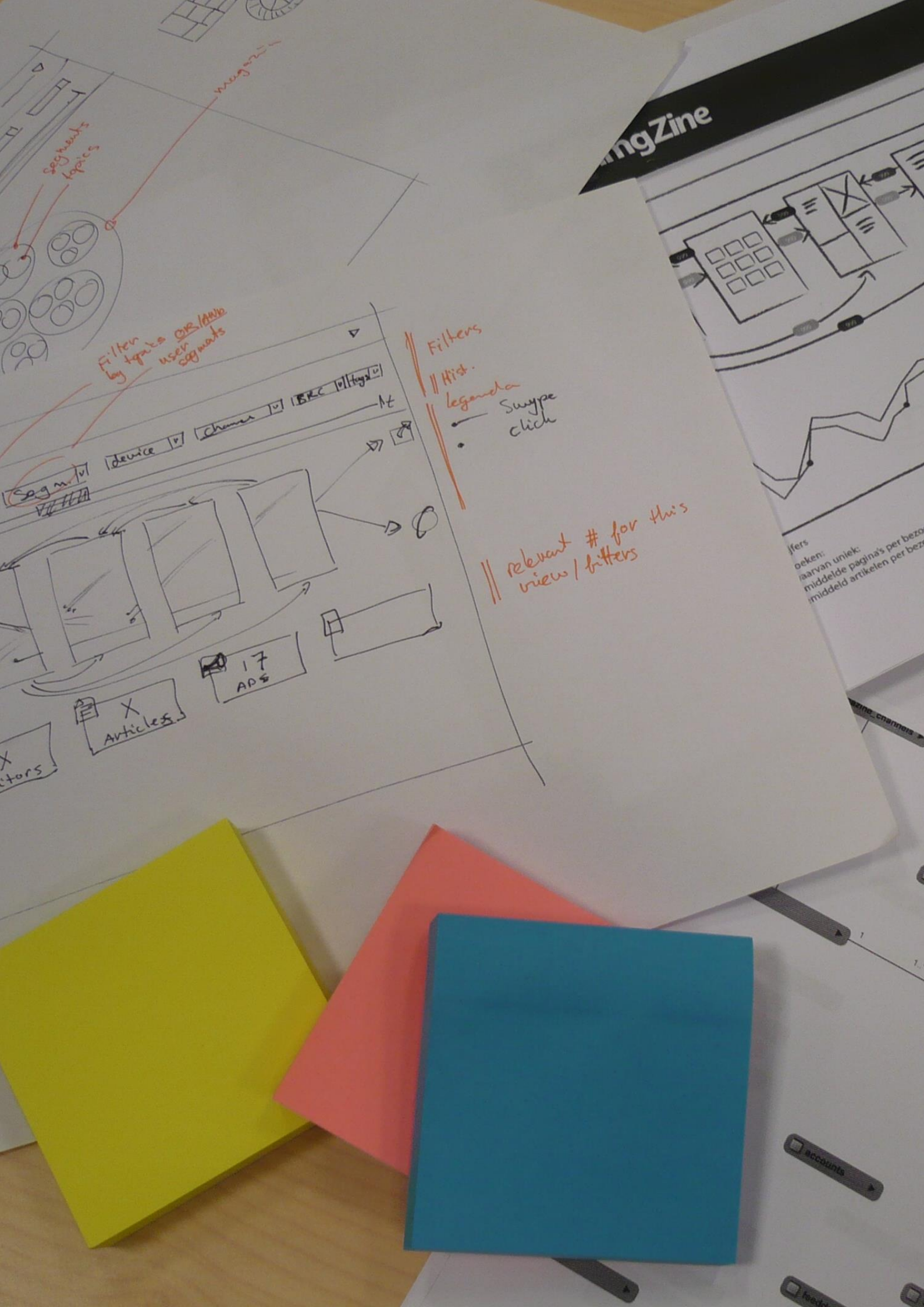
The Expert Sessions consist of a presentation or demonstration with the current status of the artifact, and some topics to be discussed about. Between one and, most often, three people join the sessions. The three themes of the sessions are Technology, Product and Sales, and Usability. Three out of four sessions are recorded.

3.5 Selection of Artifacts

A selection of five artifacts is presented in this thesis. In order to get there, the following steps are been executed. First, a fast data investigation is executed. In this investigation, knowledge about the data available, the aggregated data and measurements already developed at imgZine is retrieved. Next to this investigation, an early customer interview is executed to get a feeling of the publishers goals. Last, the current technology, the system architecture and how data is extracted was researched more extensively.

Parallel, as much as possible sketches and ideas were written down, with knowledge of the current state of technology in mind. This is done by drawing as much as possible different designs and data abstraction layers. Inspiration came from different sources, notably Beautiful Visualizations (Steele & Iliinsky, 2010), Information is Beautiful (McCandless, 2009) and the ontology found in the literature chapter (Lengler & Eppler, 2007).

Of all concepts, five are selected for more specific evaluation, specification and reachability tests. One of these selected artifacts is developed into a working prototype. First, the four artifacts that are not further developed within this thesis are discussed and after that the functional artifact.


segments
topics
Filter by topics OR/AND user segments
Filters
Hist.
Legenda
Swype
click
relevant # for this view / filters
17 ADS
X Articles
accounts

4 Artifact Description

In this chapter we describe the artifacts created, both functional and conceptual artifacts. First, we explain the different measures and technologies developed and used in-house for retrieving knowledge from data and making it measurable. Second, the selected conceptual artifacts are described from different angles: after a general description, the visualization aspects, data aspects and the results from the (internal) expert sessions and (external) customer validation sessions. Third, we dive into the development of the working prototype, based on the same format as the conceptual artifacts.

4.1 Data

Based on earlier findings, we recognize the importance of a clear link between information and the visualization used to represent it. Based on this knowledge, data is virtually grouped by shared representation to the user. By initially exploring the data gathered from the magazines, three main categories are identified: content, readers and reading behavior. These categories of data can act as view ports to the exploration of the data as seen by the user.

4.1.1 Content

Under *content* we understand the articles, including images, video's and other media and the channels in which they are grouped, which are (planned) to be published using the magazines. This articles mentioned are imported from other sources including the own Wordpress-installation and external sources like RSS-feeds or scraped from a website. Both the source and a cleaned version of the articles is stored in the database.

Additional abstractions and technologies that are also classified in this category are the topic matching algorithms. The topic matching, assigns relevant topics to articles by analyzing the text of the article and the texts related to the topics. The texts related to the topics come from Wikipedia.

4.1.2 Readers

Under *readers*, we understand information about the individuals who read the magazines and are the end-users of the delivered magazines. This is primary the information about who users say they are, e.g. by leaving an e-mail address, or selecting their categories of interest. This category of data is and will be enriched by interpreting data from reading behavior.

4.1.3 Reading Behavior

Under *reading behavior*, we understand all data, actions and other characteristics that are related to what the reader does when reading the magazines. Examples are the sequences of articles read and the time spend on reading. This is the most important group for our artifacts.

This is the information about what users do, e.g. by reading articles or swiping through the magazine. It includes every action a reader executes on the magazines. This group data group thus defines readers by their actions. Most information currently available, is aggregated from the events tables.

4.1.4 Actions and Tasks of Visualization

Now that we have distinction in types of data made just before, we can identify the tasks that could be executed by the users of the dashboard. In the literature we found a commonly used paradigm of tasks in visualizations: "overview, zoom, filter, details-on-demand, relate, history and extract" (Shneiderman, 1996).

Based on the data distinction, we identify the following six types of actions. Each of these actions can be seen as a gate to the tasks of exploration. In the final dashboards, not all these actions have to be present as individual actions. In practice, this can be used as a check list, in order to identify gaps in the data exploration tasks. The six identified types of actions are:

- content by readers;
- content by reading behavior;
- readers by reading behavior;
- readers by content;
- reading behavior by readers;
- reading behavior by content.

4.1.5 Metric: User Engagement Index

In order to distinguish articles and readers by their quality or behavior, a combined metric is developed. Simplified, the User Engagement Index (UEI) represents the reading speed of a user or a group of users, corrected for the length of the article. The UEI enables imgZine to turn reading speed into a comparable property, by including the length of the article and other relevant aspects into the metric.

The User Engagement Index is designed with the goal in mind to create insight into the reading behavior with regard to the content. It does explicitly not provide information about the opinion of the user with regard to (a set of) articles. (imgZine, 2013b)

4.1.6 Overview of Available Measurements

Using the developed technologies, five technical prototypes were already developed at imgZine. A small overview of these technologies follows:

- the **article topic graph** shows graph of the topics related to a certain article in the database;
- the **topic search** allows to search and compare the UEI of multiple topics, by plotting them over-time in one graph;
- the **full text search**: searches all texts in a magazine;
- the **viewer segmentation** distinguishes groups of readers automatically based on their behavior. Readers are grouped by their similar attributes and then presented with their group characteristics next to them;
- the **regressions** explain how big the impact is of certain factors on the engagement of the articles.

4.1.7 System Architecture

In general, performance and data structure questions are out of the scope of this research. A couple of general choices however have been made and should be noted here. As a general architecture for the analytics dashboard, the paradigm Extract, Transform, Load (ETL) has been chosen. The main argument behind this is to improve the speed on the interface. During initial development this procedure was not used yet and the general speed was considered very slow.

From the magazine events and other data is stored in the inbox, from where it is processed and the data is stored into a MySQL-database. From there, the data is in batches processed and stored again in data warehouse tables in the same database. In this processing, transformations like sums and averages are calculated. From these data warehouse tables, most information is loaded and combined for display in the analytics dashboard. In the new dashboard, more and more information is calculated near-line or even online, in order to be able to retrieve data faster. These tables are not explicitly labelled data warehouse anymore.

4.2 Conceptual: Swipe Patterns

The Swipe Patterns artifact represents the way users browse through the magazine, by representing their swipe actions across pages. Its goal is to disclose patterns in the reading behavior of users.

4.2.1 Visualization

The Swipe Patterns are inspired by the Flight Patterns visualization, see Figure 4.1 (Koblin & Klump, 2010). This is a visualization which shows the air traffic above the USA, plotting the flight routes converging to air ports.

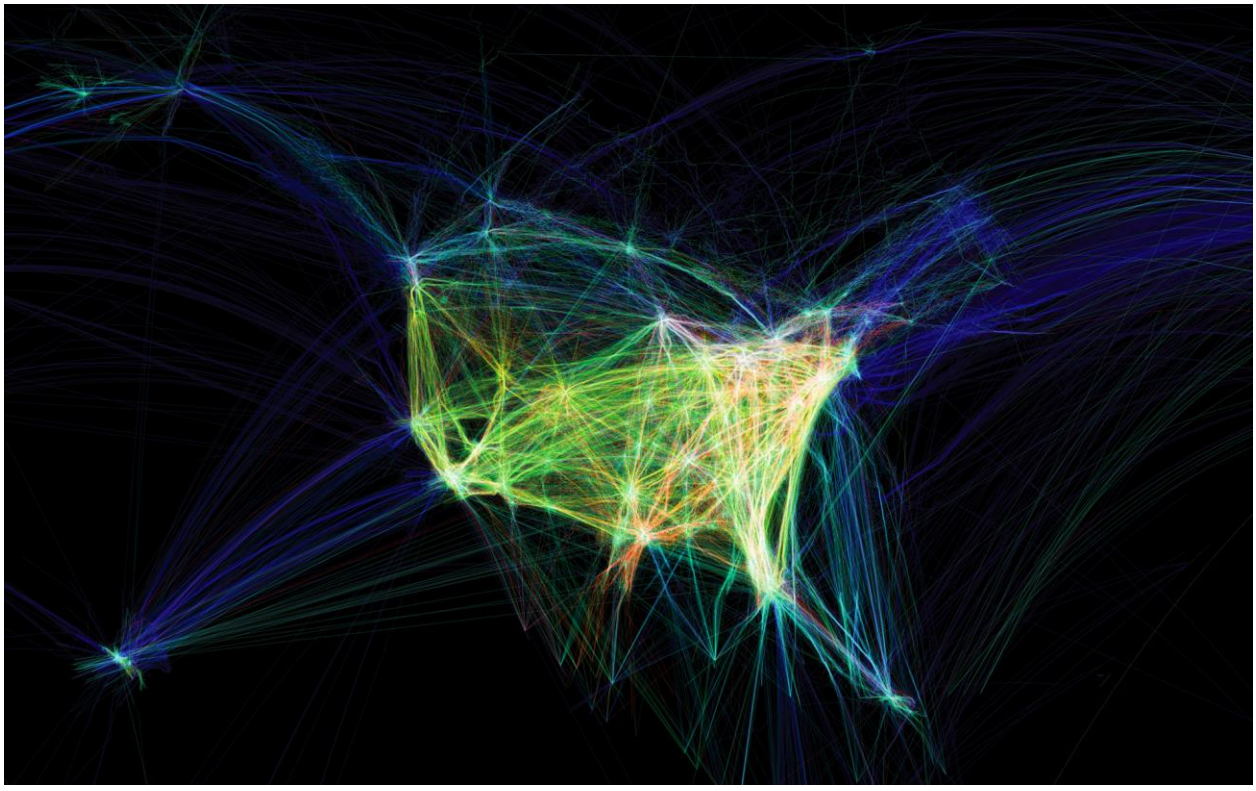


Figure 4.1: Emerging collective behavior in Flight Patterns (Koblin & Klump, 2010)

Inspired by these visualizations, the hypothesis arises that seemingly random swipes will collectively reveal patterns and insight into the usage of the application. Sub-hypotheses are, for example, the likeliness that people will not swipe on the text or place of the screen where they are reading or the part that they are using. Next to the Flight Patterns, it also builds upon the theory that visualization of data should reassemble the original form factor in a recognizable way.

Figure 4.2 and Figure 4.3 conceptually demonstrate what such patterns can look like. In Figure 4.2, it is demonstrated how swipe patterns could show behavior cross-pages, uncovering fast swiping through different articles. In Figure 4.3, it is demonstrated how swiping could demonstrate how an article is read.

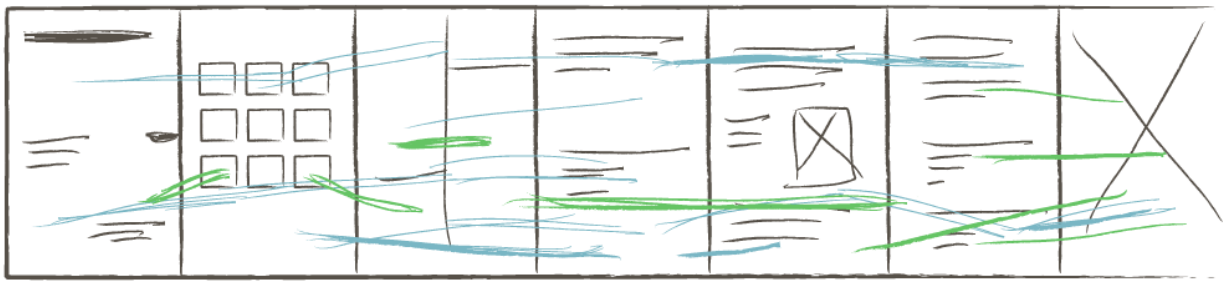


Figure 4.2: Concept of Swipe Patterns on different page types

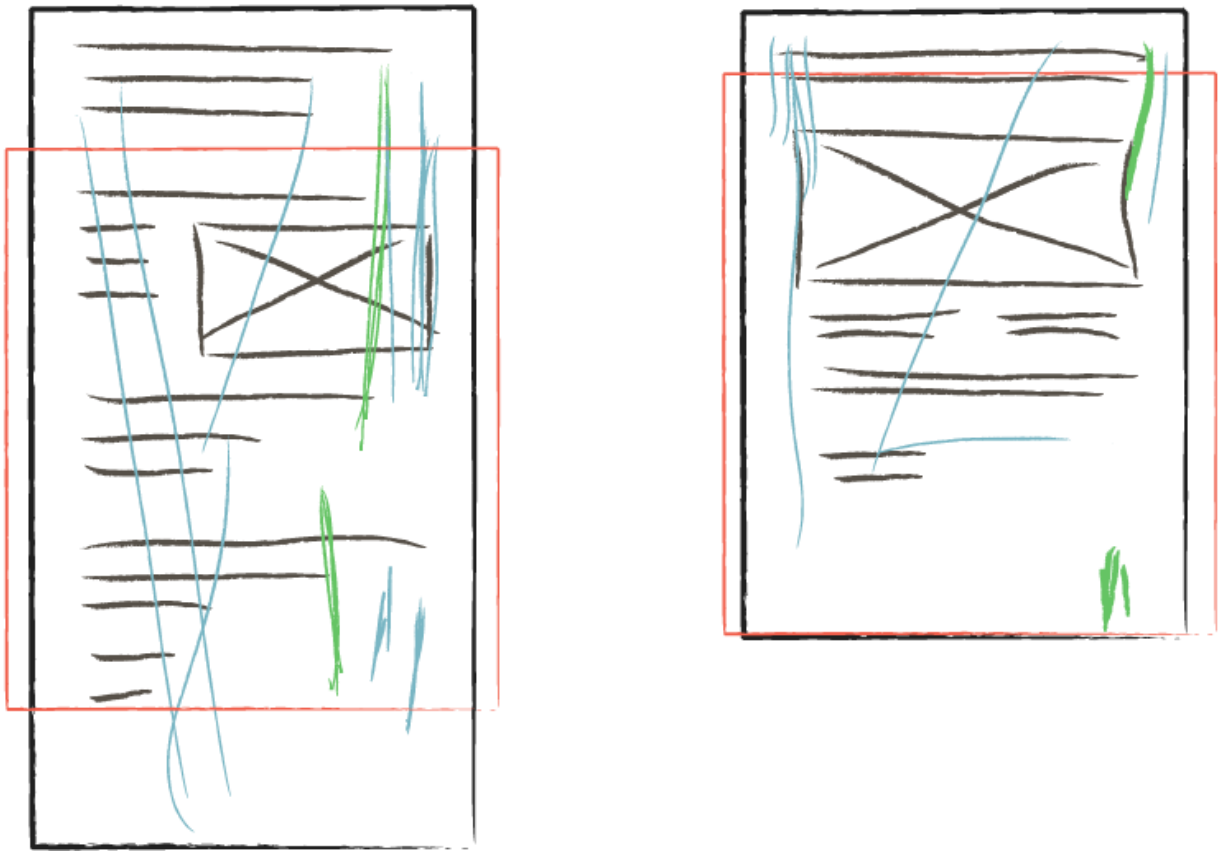


Figure 4.3: Concept of Swipe Patterns, used for comparing different articles

In these examples a distinction is made between reads and swipes. This distinction is made by using the User Engagement Index. Articles that are viewed very short, are considered swipes, whereas articles that are viewed slowly are considered reads. Extreme cases are filtered out.

4.2.2 Data

For tracking the swipes, data regarding the start- and end-positions and the specifically used templates (in the 'magazine'-views) are needed. Currently, these templates are picked randomly from a predefined set. In new built technology, the templates are picked based on the characteristics of the articles that are involved.

Currently, only data regarding events, for example from page to page is available. Due to the dynamic templates, it was not easy to identify and store which exact templates are used. However, during this research, a solutions for storing the templates is designed. It is done by storing the coordinates of the boxes, which can

later be combined with the articles. This data was not available in the conceptual phase of this research, and therefore this artifact was not selected for functional development.

4.3 Conceptual: Enriched View Trail

The Enriched View Trail gives insight in the aggregated user flows of readers, when using the magazines. Its goal is to give an overview of and changes in, the usage of the magazines. The colors representing changes in the data, but could also be used to represent the differences between two magazines, when comparing them.

4.3.1 Visualization

This visualization used as basis is the view trail, currently present at the imgZine customer analytics dashboard. It does display the click through rates in the magazine and is based on the structure of the magazines created by imgZine. In Figure 4.4, the current situation is displayed. Not all data is present, as can also be seen in this figure. This is not rarely due to an implementation error at the earlier magazines. Some customers have an dashboard in which the view trail is not present.

In Figure 4.5 the enriched view trail is displayed. Its main improvement is the addition of color coding to the numbers, showing increases or decreases in the number of visitors. This way, the figures can be put into perspective making it easier to compare the numbers (McCandless, 2010).

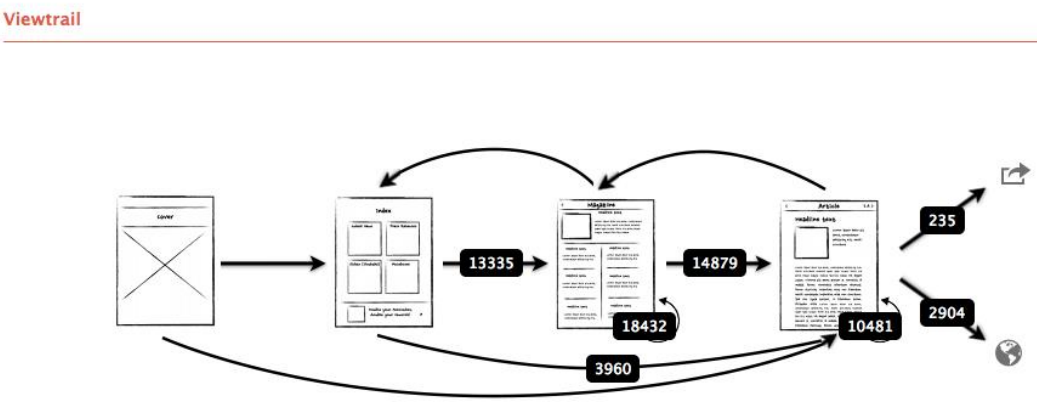


Figure 4.4: View trail in the current Analytics Dashboard

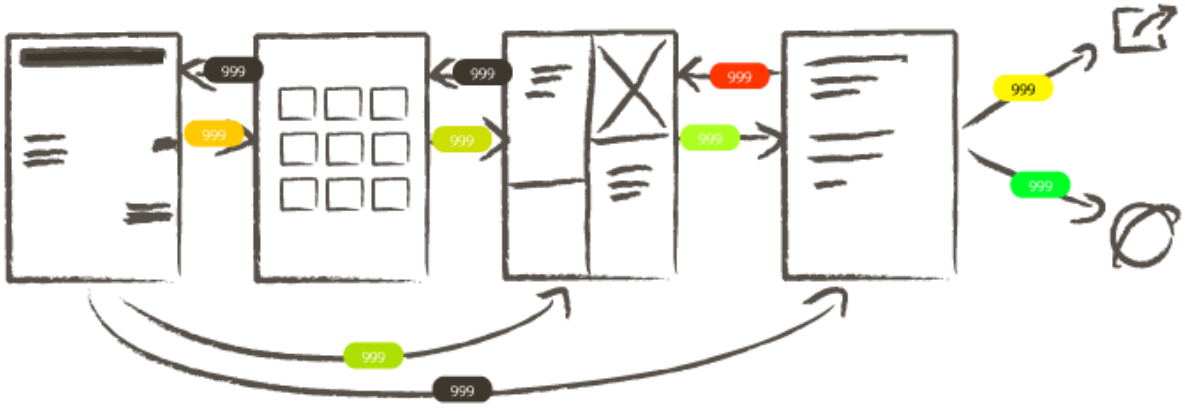


Figure 4.5: Enriched view trail

4.3.2 Data

The data needed for this visualization is available. The relevant events are currently logged, starting with an appstart-event, when the app is launched and ending with an appclose-event when the app is closed, or pushed to the background. In some magazines events on the cover page are not fired, while on some devices the appclose-event is not always fired. This is however corrected. In the current view trail, the number are aggregated in a data warehouse (table) before they are used in the graph. When calculating the values for the color coding, interval calculations need to be made from the data. As it is likely that no extreme performance issues will occur and thus the currently present data can be used.

4.4 Conceptual: Topic Bubbles

The Topic Bubbles artifact is created to improve insight in what content is published and how popular it is. Its main goal is to create an overview of the published content, which can give a complete overview of the content and its current popularity in the blink of an eye. It can be used to for example compare channels or magazines against each other.

4.4.1 Visualization

The Topic Bubbles artifact positions collections of articles on fixed locations, in order to compare multiple magazines, channels or other sets of content among each other. The fixation of positions is done by allocating topics to articles and plotting the articles based on their associated topic(s). Thus, the topics actually have fixed positions in the circle, based on their subject. This can be used to compare different content, but also to compare the content *published*, with the content *read*. On top of the topic circles about what is published, an overlay of the most intensively read articles can be plot.

The visualization can be seen as a mapping of a 3D graph network on a 2D space, in order to make it easier to present and compare.

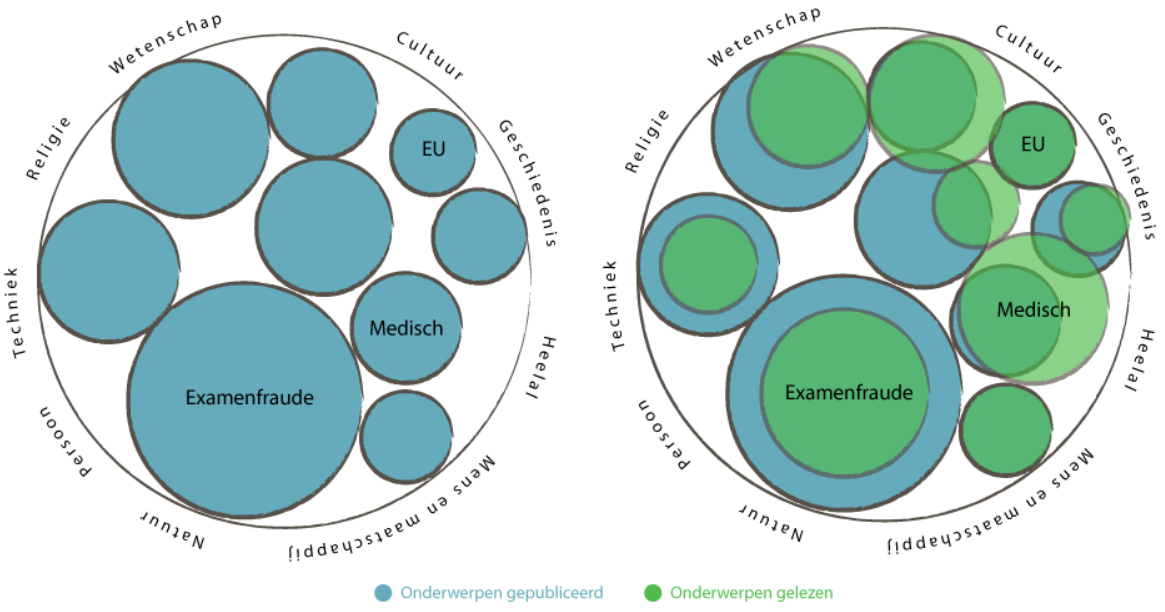


Figure 4.6: Concept of Topic Bubbles. Left without comparison data, right with User Engagement-data.

4.4.2 Data

To display articles like this, the content is matched to topics by natural language processing (NLP) technology. Topic categories and their meanings are based on data retrieved from Wikipedia in the relevant language. This topology is combined with a more hierarchical set of topics to create a topic hierarchy. The size of the bubbles represents the amount of articles published.

To test the hypotheses, 23 data analysis studies have been carried out, most of them using an open source visual data mining tool called Orange (Bioinformatics Laboratory, 2013).

Using Orange a set of analytics are executed to look for comparable sets of information present in the content of the magazine. This was a rough search and most results confirm the design of the magazine and backend, making them self-evident. Others are giving directions that such visualizations could not be made using the currently used algorithms.

An example of such a explorative calculation is shown in Figure 4.7. This figure teaches us that (it is likely that) the topics are evenly distributed across people that view articles fast (swipers) and people that view them for a longer time (readers). Extreme cases also seem evenly distributed over topics.

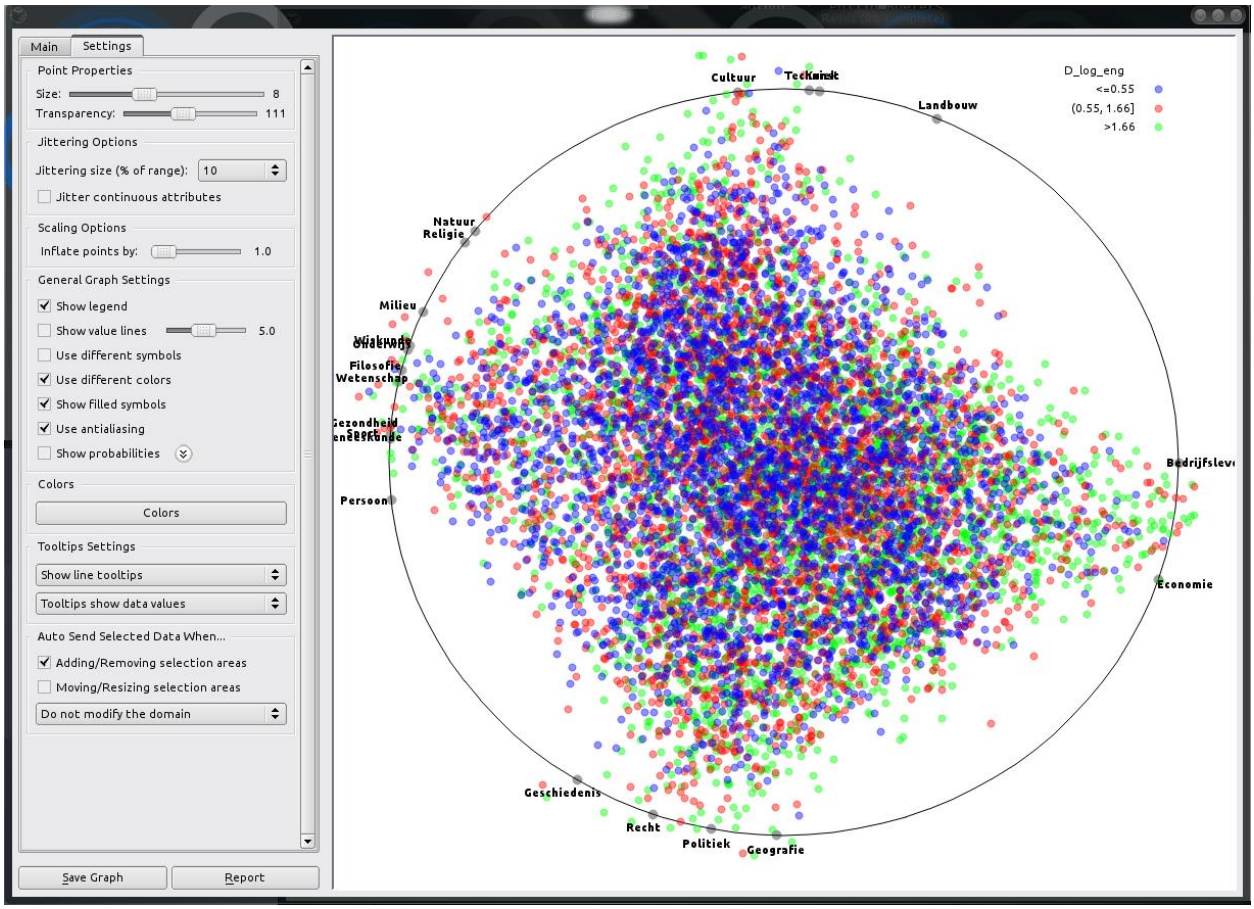


Figure 4.7: Analysis of the reading intensiveness across different topics using Orange

4.5 Conceptual: BubbleUp

The BubbleUp artifact reveals the topics/articles people are the most engaged with. It can function as a real time visualization of the content published, or in order to distinguish which topics are (not) worth publishing about.

4.5.1 Visualization

The visualization, seen in Figure 4.8, is based on the 'Snake Oil?' visualization (McCandless, 2009, pp. 18–19), which distinguishes dietary supplements by scientific evidence. Topics that have articles with higher user engagement indices, bubble up to the top and other fall down. The horizontal line is the Worth It Line, meaning that a topic below the line, might not be worth publishing, because it is not read well enough. The value might be the average value, or another value based on intelligence.

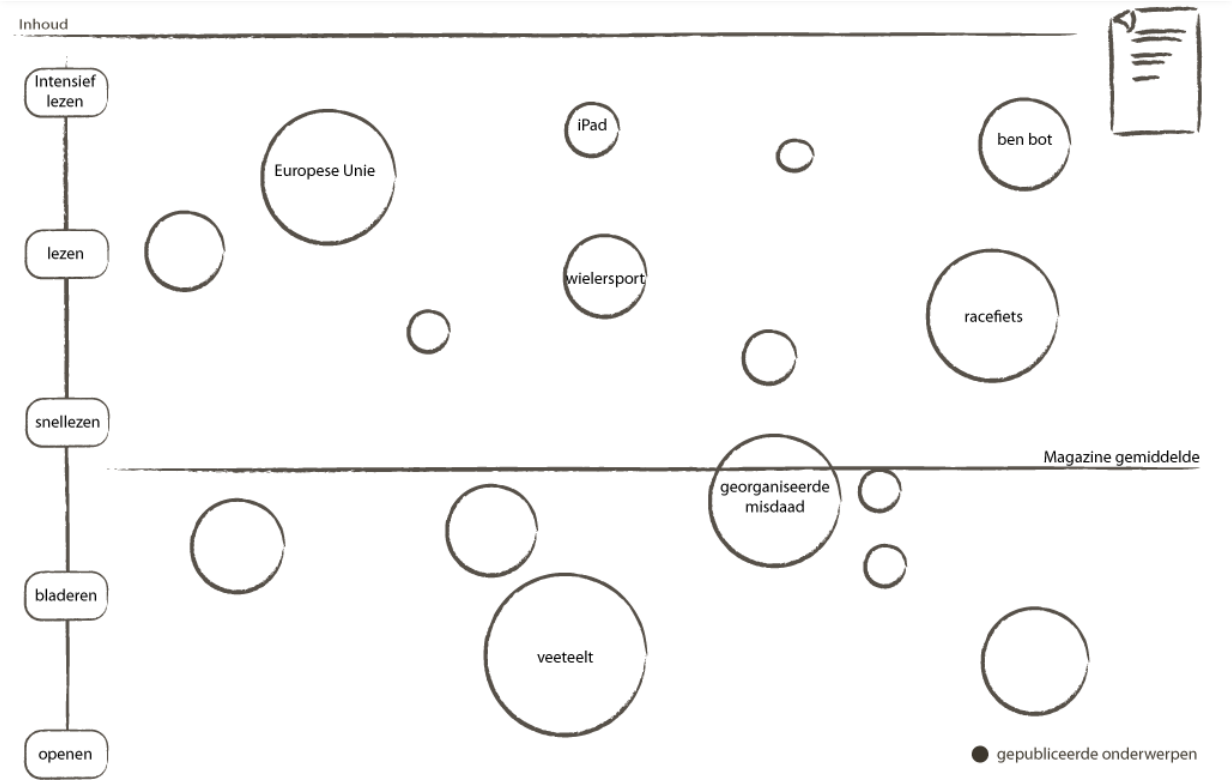


Figure 4.8: BubbleUp artifact

4.5.2 Data

This concept uses almost the same data as the Topic Bubbles artifact. The bubbles are articles grouped by topics. The average engagement value (y-axis) is calculated by a logarithmic scale, to reduce the effect of outskirts on the visualization. The initial composition of the data was easy and took less than an hour to execute, but further optimization is needed, as is explained more in-depth in section 5.6.1.

4.6 Functional: User Flows

The functional artifact User Flows gives insight into the paths that readers have followed throughout the magazine. Its goal is to represent these in an understandable and easy accessible way. When developing the artifact, the learnings from the earlier conceptual artifacts have been used. Especially the artifacts regarding reader paths, namely Swipe Patterns (section 4.2) and Enriched (section 4.3), are used as inspiration.

4.6.1 Visualization

The visualization of the User Flows artifact is based on the Enriched View Trail, and the Swipe Patterns. In contrast to the (Enriched) View Trail, the User Flows artifact includes include all article views as individual

states. In contrast to the Swipe Patterns, it does not represent specific swipes, but only an abstraction of the reading behavior is presented. In Figure 4.9 a screenshot of the visualization is presented.

It used two metaphors: the use of icons for representing the pages in the magazines, and the use of a trash for leaving the magazine. In the evaluated version, also dots are used for the more-state (not present in Figure 4.9). Not all intended visualization aspects are developed into the working artifact, due to time limitations. For example, the use of a paper stack metaphor for visualizing the number of views, and an use of screenshots of the actual magazine, instead of icons.

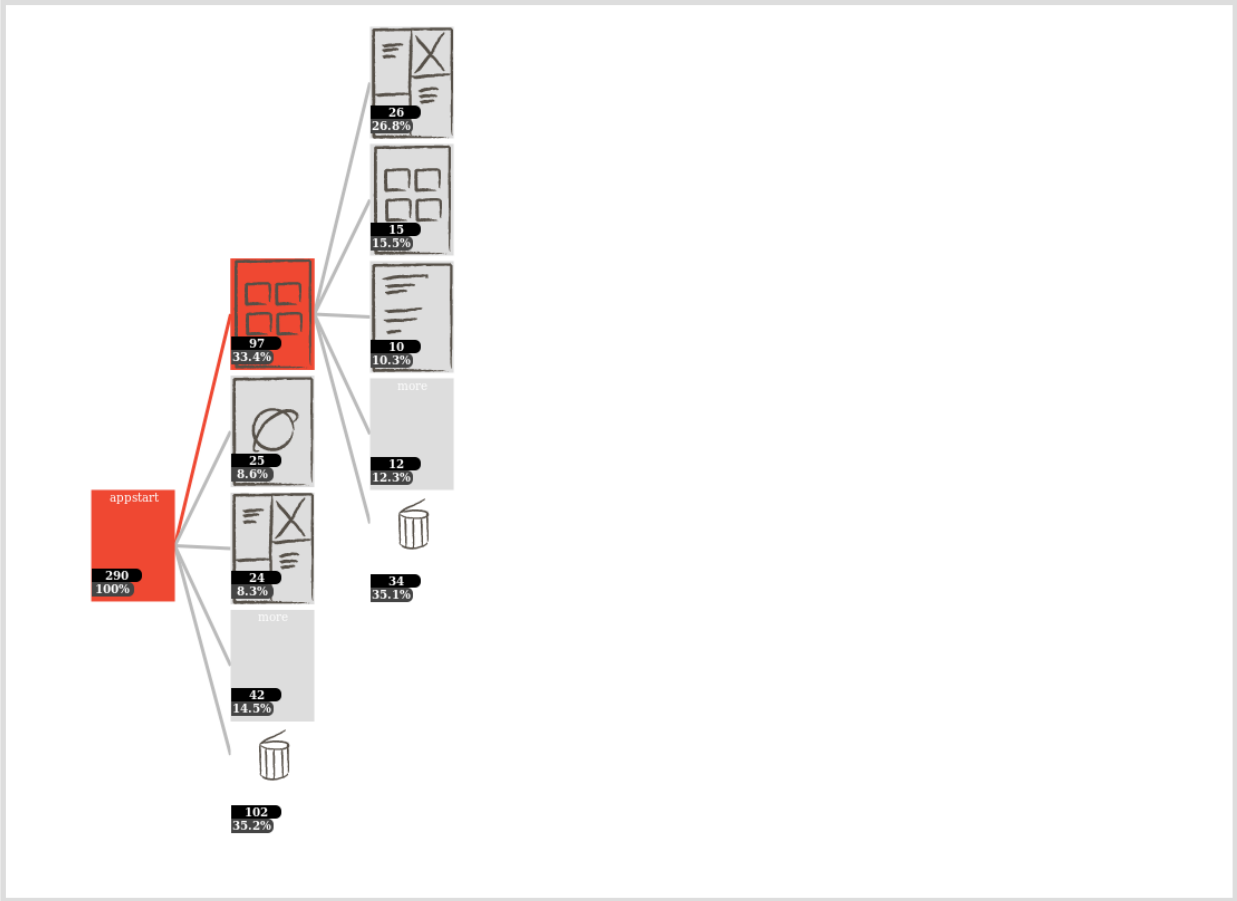


Figure 4.9: Screenshot of the functional artifact with one day of data from one magazine

4.6.2 Data

The data for the final functional artifact was created after eight iterations created for data exploration and combinations. Initial versions focused on identifying the states available and their relations in a mathematical manner (for creating an overview of the data), where later versions focused on further exploration, combination and verification of the integrity/correctness of the data. Roughly, four phases in the exploration process can be identified:

- Exploration of states and their relations;
- Exploration of unique states;
- Verification of correctness of the data;
- Extraction and combination of the data.

Each of these phases are shortly described hereafter.

First, the **exploration of states and their relations** is done using a python script which loads the event data from the database and looks for a sequence of an individual user session. Each individual session is added to a directed graph, from which a representation is generated. This way, the actions are not threaded uniquely. For example, two different articles are still considered one single state. A graphic of this representation is found in Figure 4.10.

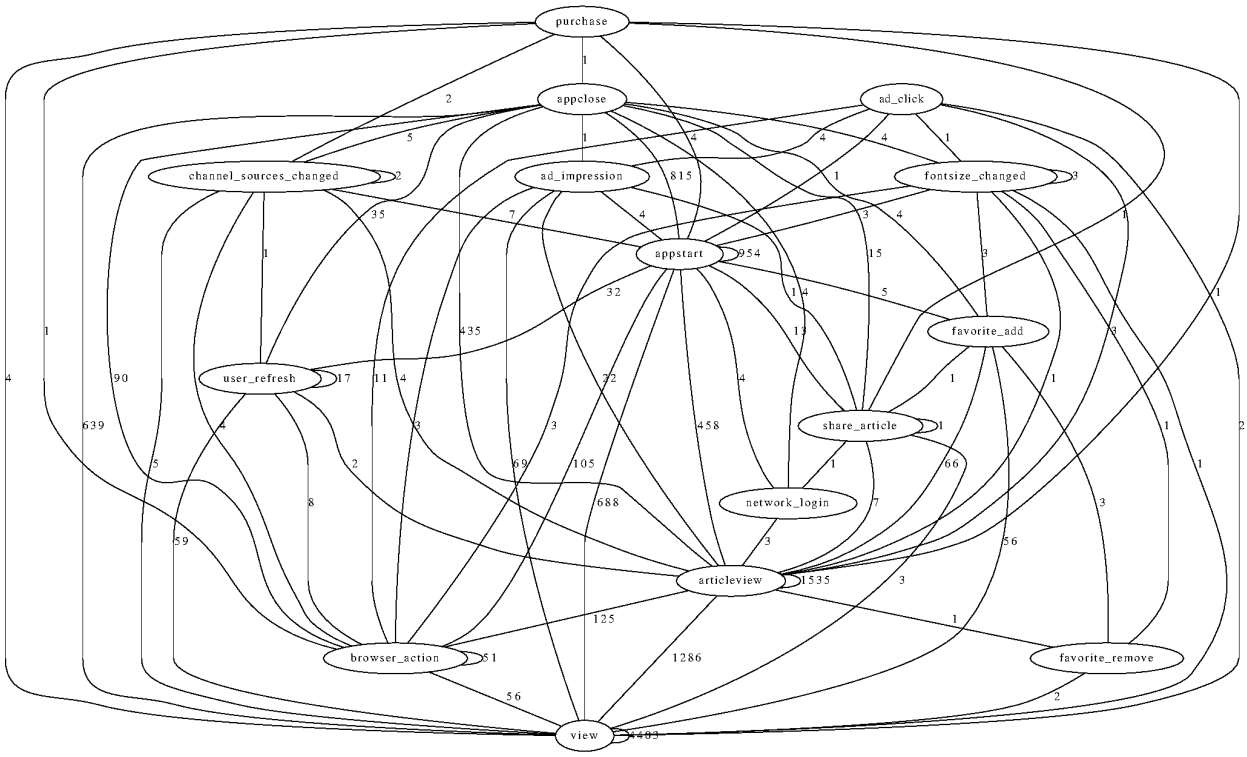


Figure 4.10: Overview of the events and their relations for a common magazine

Second, in the **exploration of unique states**, each action is treated as unique. For example, every article viewed in the magazine is treated as unique, in contrast to all articles considered the same in the previous state. The events are extracted from the database and a directed graph is created. In Figure 4.11 an example of such a graph is shown. It might become clear to the reader that this graph has a low readability and thus a better visualization would be suitable.

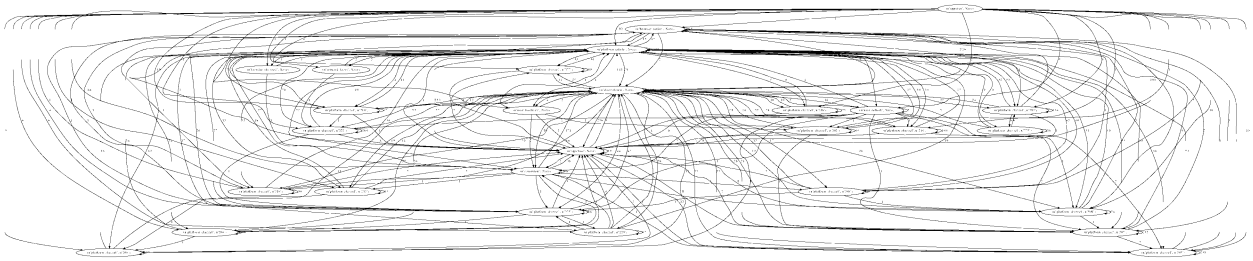


Figure 4.11: Overview of the events and their relations when using unique states

It was found that the used graph library did by design not support multiple instances of the same state (as every node is considered a state). In order to work around this, in four iterations, new prototypes are build. In

these prototypes the data tree is combined without a facilitating library and the nodes and format transformations are executed using recursive functions.

Third, **verification of correctness of the data** is executed. Using both the earlier versions of the representation build and regular database software, strange effects were identified. Every exception was discussed and validated by a colleague with a background in artificial intelligence. Two notable problems have been identified for different artifacts:

First, events before July 19, 2013 were not retrieved from the inbox with all information. As a results, the states created did contain empty field (NULLs) and many were less unique as they should be. This bug was externally to this project and has been fixed.

Second, building the tree using recursive functions creates a limitation: a set of children from a node cannot be changed while iterating over it. Initially, a work around was found, but when using larger sets of data (starting at about one month, containing more than 3000 nodes), in some cases the bug still existed. However, this is not considered a big problem when evaluating the artifact. This bug could likely be solved by using an iterative instead of a recursive approach.

Next to the explicit validation phase, each of the iterations is validated by asking an internal data mining expert to check if the correct data is extracted. Also, the sum of the total number of events in a period is made and compared with the total number of events present in the visualization.

Fourth, the **extraction and combination of the data** has been executed several times before the final version evolved. In the final version, we construct the paths for every individual device (which are often referred to as unique visitors) by sorting the event table by date and searching for the event first fired when opening a magazine (“appstart”) and add events till we find a closing event (“appclose”). In case such a path is not complete, for example, no close event is found, the path is discarded. The combination of all paths is summed and leads to a tree, which always starts with the appstart node as root.

5 Evaluation

In this chapter, the results of the evaluation of the artifacts are presented. For validating our hypotheses and by doing so answering our research question, we need to evaluate our designed artifacts empirically. For doing so, a couple of validation methods were used. First, expert sessions and customer interviews are held for validating and selecting the conceptual artifacts out of a set of 56 drafts and ideas. The customer interviews are also used to get a general view of the perspective and wishes of the customers. Second, the functional artifact is also validated by a couple of internal and customer interviews.

5.1 Types of Customers

Based on the customer interviews, four possible types of customers, over two axes are identified, as being the most important distinctions, as can be seen in Figure 5.1. This figure is based on the customer interviews and an small internal discussion for validation. The positions and regions are subjective interpretations, to get a general idea and overview of the types of customers and their goals. The positions are based upon the part of the interview with general questions.

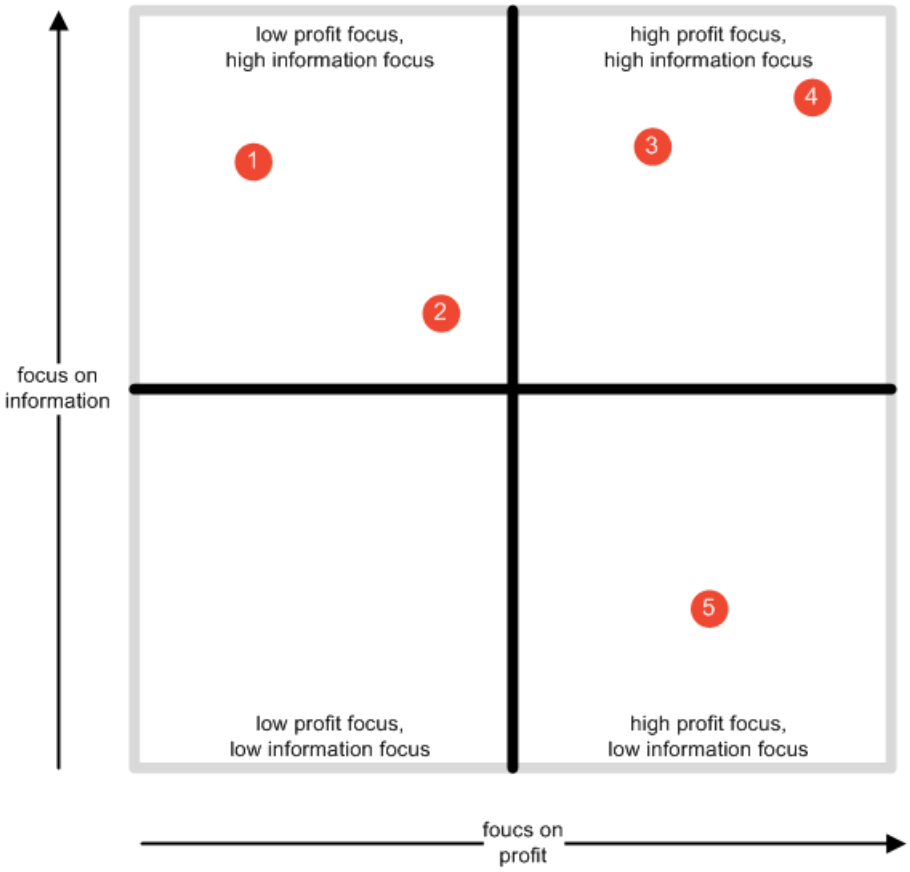


Figure 5.1: Overview of customers and their focus. Number 1 and 2 are regarded as enterprise publishers; 3, 4 and 5 as traditional publishers.

On one axis, the customers are identified as mainly publishing for profit making (most likely external magazines) or mainly for information delivery (most likely internal magazines). Under 'focus on profit' we understand to which extent the customer gave answers to questions that are directly related to optimizing their (primary or otherwise) cash flows. On the other axis, the amount of involvement with the further exploration of information from the magazine, the analytics, is used. Under 'focus on information' we

understand to which extend the customers gave answers that were expressing a wish for more or further information about how and why the magazine is used.

The four quadrants are defined to get a general distinction of the types of customers and their goals. They are shortly described hereafter.

Customers in the upper left quadrant want to deliver content that is read very well, and does transfer their message in a positive way. They want to understand their readers, in order to **deliver better content** in a better way to them. There is little to no incentive to directly generate revenue from the magazines. Both customers in this quadrant, 1 and 2, are enterprise publishers.

Customers in the upper right quadrant want to deliver better content, while selling more content. They want to understand their customers to be able to deliver better content and by doing so, sell more. Revenue is mainly generated by **selling advertisements**. Both customers in this quadrant, 3 and 4, are traditional publishers.

Customers in the bottom right quadrant focus on selling more content. There is a high incentive to generate revenue. The customers in this quadrant wants to **sell more subscriptions** in order to stay in business. The customer in this quadrant, 5, is a traditional publishers.

5.2 Conceptual Artifacts

Using the expert sessions the feasibility and reachability of the 56 drafts and ideas were tested and five conceptual artifacts were selected for a further evaluation using expert sessions and customer interviews.

The expert sessions were organized with employees (full-time, or contracted) of imgZine, who were aware of most characteristics of the products. In the *Expert Session Technology*, three people joined, who are involved in delivering new technology for the magazines and the analytics dashboard, for example the metrics and topic mapping technology explained earlier. In the *Expert Session Product*, three people joined, who are involved in designing and coding the final magazines delivered to the customer. In the *Expert Session Usability*, two people joined, who are concerned with the usability and general look and feel of magazines.

The customer interviews are executed under the five customers described in the previous subchapter. Based on these interviews, Figure 5.1 was created, containing an overview. Customer 4 was interviewed earlier in the process, as an orientation. The other customers were presented the same materials, although different selections of artifacts were made, based on their understanding of the concepts. Of the five customers of imgZine that are involved in the interviews, two are regarded as enterprise customers. As we found from the first round of customer interviews, these tend to have a higher level of information expertise and thus are asked for the evaluation round of the final functional artifact.

5.3 Conceptual: Swipe Patterns

In this section the evaluation of the conceptual artifact Swipe Patterns is described. A description of the visualization, data and other characteristics of this artifact can be found in chapter 4.2.

5.3.1 Internal Evaluation

The Expert Session Product and Sales notes the Swipe Patterns are an interesting idea, but that heat maps might already do the job. They are not sure if the results will be directly useful, but do expect them to be at least interesting or useful as a cool demonstration. For this representation, the swipes should be mapped on the actual layout, so templates should be numbered and tracked. Also the start, end and duration of a swipe

has to be recorded, which is quite some data together. The idea of tracking swipes might be interesting for article views, e.g. on how to improve the written article.

The Expert Session Technology is less positive, but does see a role in representing on an individual article level. They note that it might very well be possible to track swipes in order to deliver the data for this prototype. Currently, the data is not available and some updates on the core components are needed in order to realize this. It is also noted that this would be useful in combination with the gyroscope for example, in order to identify the orientation or other ergonomic aspects of the device usage.

5.3.2 Customer Feedback

Several customers noted this artifact could well give insight in the usage of the magazines. Especially for particular articles and layout of magazine templates, they expressed interest in the way people are reading these articles. Those regarded as traditional publishers generally expressed themselves more positively than those who are regarded as enterprise customers. Publisher 3 noted that “people are not reading text, but layout. Design and positioning is far more relevant as most people think”. For this matter, they are interested in improving the layout of individual articles or pages.

One traditional publishers notes that it is important to distinguish different types of devices.

5.4 Conceptual: Enriched View Trail

In this section the evaluation of the conceptual artifact Enriched View Trail is described. A description of the visualization, data and other characteristics of this artifact can be found in chapter 4.3.

5.4.1 Internal Evaluation

The Expert Session Product and Sales sees the view trail as a start for further analysis and intelligence in the dashboard and agrees it could be further extended.

The Expert Session Technology likes the idea to extent the view trail, as they find it a clear foundation. It is noted that it is a difficult decision which data exactly will be used when showing increase/decreases.

5.4.2 Customer Feedback

Interviewees are generally positive about the view trial, but some need hints toward the magazine structure. Not every customer is aware of the structure of the magazines and thus it is not always directly clear which type of page is represented. Two customers note that it would be interesting to see how many people click through after the first overview. One customer expresses the explicit wish to dive deeper into the data, with the view trail as a starting point.

5.5 Conceptual: Topic Bubbles

In this section the evaluation of the conceptual artifact Topic Bubbles is described. A description of the visualization, data and other characteristics of this artifact can be found in chapter 4.4.

5.5.1 Internal Evaluation

The Expert Session Product and Sales regarded the concept of topic bubbles as very interesting. If it would be possible to create a 2D overview of the content, it would allow easy comparison between content published and consumed. An idea that appealed the attendees. On the other hand, both in the Expert Session Product and Sales and in Usability, it is noted that the concept is abstract and needs further development leading to

clarification and simplification in order to be useful, or even explainable, to customers. It is also highlighted that using real world data would greatly improve the communication of the concept.

The Expert Session Technology was less positive about the concept. In order to create insight, the members regarded the identifications of groups of similar articles/topics more important, than fixed positions of the topics. A difficult point is the mapping of topics on some sort of ontology/categorization of topics. The idea is to capture the topics as bubbles in a circle, so they have a common point they are matched on. Most common visualization of articles and user profiles are bubbles currently, but they do not capture hierarchy. The Wikipedia category structure might be used, as Wikipedia has about 9 top-categories, although there are sub-categories that are cross-mapped.

5.5.2 Customer Feedback

The concept is relatively abstract and it is hard to communicate it without real world data. In the current state, the concept was not positively received by the customers in the Customer Reviews. Reasons mentioned included too complex, or too abstract. One customer mentioned traditional bar charts as an alternative.

5.6 Conceptual: BubbleUp

In this section the evaluation of the conceptual artifact BubbleUp is described. A description of the visualization, data and other characteristics of this artifact can be found in chapter 4.5.

5.6.1 Internal Evaluation

From this conceptual artifact, a working prototype is created, using a set of real data. Due to a server crash this prototype was lost before it could have been evaluated as a working model. The concept gave some insight and revealed, through internal discussions, the following problems.

First, the topic bubbles are created by the number of articles in a topic, but some form of degradation of data is needed to represent a correct size for the topic bubbles. Some topic could have been a great discussion in the past, but that does not automatically mean it is currently still important, if one recent article is published about the topic.

Second, it is hard to decide how the information on the x-axis should be grouped. The graphic used for inspiration has subjective grouping by related bubbles. A combination of time-based degradation and topic-relating could be a further direction.

5.6.2 Customer Feedback

To some customers the concept was not exactly clear, while others found it interesting. One customers notes the information interesting and useful, but the representation is more complex as needed. A bar graph would be sufficient to display this. Another customers notes it could give an overview of the content, although they do not expect to use it.

5.7 Functional: User Flows

In this section the evaluation of the functional artifact ‘User Flow’ is described. A description of the visualization, data and other characteristics of this artifact can be found in chapter 4.6. For evaluating the functional artifact, two customer interviews were executed, with customers selected based on the matrix earlier presented in Figure 5.1. The relevant customers are number 1 and 2 in this diagram.

The selection is based on the following criteria: the customers are highly involved in analytics and they are do not focus on directly generating revenue from the magazines. These customers also tend to be enterprise

customers and thus have, based on the first interview, more expert knowledge in the field of information management.

The semi-structured interviews are carried out, while demonstrating the functional artifact in an interactive manner. The interviewee is asked to click around in the artifact and discussing its usefulness. To support the process of exploration and validation, open-ended questions about the artifact were used for guidance.

5.7.1 Customer 1: Belastingdienst

The magazine produced for customer 1, the Belastingdienst (Dutch Tax Agency), is at the phase of being transferred from pilot to daily use. It is used for disclosing the intranet with news and other updates about the Dutch Tax Agency to the employees. The pilot is considered successful by the interviewee. The interviewee works for the Team Innovation and Communication, responsible for new technologies, like the magazine app. She is involved in the development and deployment of the magazine from the start and is highly interested in anything that could improve the user experience in order to improve the usage. She has experience with the old imgZine dashboard.

The interviewee directly recognizes what the artifact is intended to represent and she identifies most icons as intended. She would use it, but does not want to click around freely by herself. A suggestion proposed is to identify the most viewed, or longest-read path.

A few specific improvements are put forward: the flows should be more explicitly visualized, for example by adjusting the thickness of the lines and nodes should represent more information, for example give access to the User Engagement and mention which specific article is presented.

The interviewee further points out that the comparison of different periods and times could be very interesting for them.

5.7.2 Customer 2: Rabobank

The magazine produced for customer 2, the Rabobank, a Dutch co-operative bank that is internationally active, is in daily use. It is intended for Dutch consumers/entrepreneurs who would like to have information about specific sectors. It is promoted to this public by radio and television broadcastings and through other channels. The channels in the magazine are organized according to different industries and content consists of combined news feeds and specifically written blogs for this purpose. A special feature/widget is the ‘figures and trends’, which represents information from a database containing (mainly) industry figures. The interviewee works as information specialist. He has some experience with the old imgZine dashboard. Next to that, the interviewee receives on request an Excel file with all articles, the number of views and some extra information, something that is only partially available on the old dashboard.

The interviewee recognizes the function of the artifact quickly and identifies most icons as intended, except for the ‘more’-node, which is seen as a scrolling action.

He sees the artifacts as most suitable when considering a redesign of the application. He would thus probably not use the artifact as that is currently not his focus. He only sees a use for the artifact if it contains more context, e.g. it could be useful if it says something about specific articles. The main reason behind this is that the blogs produced are really time-consuming and they are thus interested in how often, by whom, why and how they are read.

The interviewee sees some illogical paths and would like to have more context about them and for example make the distinction whether or not a reader returned, or the session was a new one.

The interviewee uses his iPad with the magazine during the session in order to identify the links between the magazine and the artifact.

5.8 General Results

From the customer interviews, some general results are identified next to the results above. These general results are not allocated to a specific artifact.

First, one customer expresses that she misses information on the interaction with the app. The same is noted as interesting by another customer, but they would use it primarily for their analysts in order to improve the magazine in general and not use it themselves.

One customer expresses the wish to be able to share information from the dashboard, as it is currently copied and combined by hand.


BlackBerry

6 Conclusions and Discussion

In this chapter, the evaluation of the artifacts is discussed, along with the creation of the artifacts. The first three sections represent the pillars used throughout this research. For each pillar, the corresponding sub research question is answered. In section 1.1, general conclusions are drawn. In section 6.5 the limitations and threats to validity are discussed, after which the research contributions and recommendations for future work are presented.

6.1 Big Data

In this section, we discuss and draw conclusions based on our first sub research question. This first sub question concerns the research pillar Big Data. First, let us reiterate the question:

SQ1: ***What*** *data about content, readers and reading behavior is already available, or can be (easily) acquired, from the imgZine magazine platform and how can we extract this data?*

6.1.1 Why this Question?

In order to be able to determine which information can or should be represented, and to design visualizations, it is important to know which data is available. Also, in order to be able to use the data, the ability to extract, modify and aggregate it should be tested.

6.1.2 How is it Answered?

The process of data handling is split into two phases, the Data Exploration and Data Composition phases. As exploring and extracting the data is part of the process of building the artifacts, both phases are described in chapter 1.

In the Data Exploration phase, three categories of data that are relevant for analyzing and processing are identified. *Content* is all media content that is (about to be) published in the magazines, the *readers* are the users and who they say they are and *reading behavior* is what they do. The five artifacts especially focus on representing *content* and *reading behavior*.

In the Data Composition phase the categories identified earlier are further investigated and extraction is tested by developing prototypes. Only *content* and *reading behavior* data is examined, as these categories of data are needed for the artifacts. For creating the Topic Bubbles and BubbleUp artifacts, primarily content data is analyzed. Articles are grouped by topics in order to reduce the data presented. The User Engagement Index is used to distinguish articles that are read extensively from those that are not. For creating the Functional User Flows artifact, individual reader paths are aggregated from raw event data. These paths are accumulated, and a data tree is created.

6.1.3 What are the Findings and Conclusions?

The data used at imgZine can be allocated to the first and second generations of Business Intelligence & Analytics (H. Chen et al., 2012). The event data and cleaned content data is BI&A 1.0. It is stored in a MySQL database and therefore easily accessible and extractable. Also, extracted and aggregated data, like the topics taxonomy, belongs to this category, as it has been modified from unstructured web-based data toward structured database tables.

The raw content loaded is, in many cases, scraped from websites. Therefore, it is web-based/unstructured data, and can be regarded as BI&A 2.0 data. The social features of the magazines, for example a Facebook or

LinkedIn account that can be linked to the content sources, also belong to this category of data, as the text and images are also unstructured data.

Currently, no typical mobile/sensor-data has been used as part of this research, and no Location Based Services are in development. There is thus no data in the category of BI&A 3.0 involved.

6.2 Visualization

In this section, we discuss and draw conclusions based on our second sub research question. This second sub question concerns the research pillar Visualization. First, let us reiterate the question:

SQ2: ***How*** *can we visualize this data in an easily understandable and accessible way?*

6.2.1 Why this Question?

Visualization is an important facilitator in communicating. (Hibbard, 2004). Earlier we stated in the problem cause, that it is difficult to create intuitive access to multi-dimensional data with traditional data/information visualizations, while at the same time the research motivation states that digital publishing technology enables publishers to better understand their content, readers and readers' behavior. Currently, publishers are not skilled or trained to use the digital publishing tools they need. Out of the alternatives, education is an expensive and on-going process, data visualizations are complex, and currently known information visualizations are overwhelming.

6.2.2 How is it Answered?

For each identified content type, one or more artifacts are created in order to create insight. From a total of 56 drafts, five conceptual artifacts were selected in Expert Sessions and Customers Interviews. One Functional Artifact has been developed, which was also validated in an Expert Session and Customers Interviews. The visualizations of the artifacts, and what is learned from it, are described in more detail in chapter 1.

6.2.3 What are the Findings and Conclusions?

There are unlimited ways of visualizing data or information; therefore the best one can never be found. The five visualizations that are developed as artifacts, are shortly discussed hereafter. The findings and conclusions here focus on the visualization and the findings while developing them. The Customer Evaluation is integrated in section 6.3.

The **Swipe Patterns** artifact, described in section 4.2.1 , shows how people swipe through the magazine by visually plotting their swipes. It reveals usage patterns, and it is expected that collective patterns emerge. The artifact might give insight in where people focus on when reading. However, this has not been (in)validated yet. The artifact also gives insight into the percentage of the article that is read. This way it can be used to identify obstacles like images or badly written paragraphs.

The **Enriched View Trail** artifact, described in section 4.3, shows on a generic level how users behave throughout the magazine. Its main enhancement over the currently used View Trail is in adding color coding for increases or decreases in the number of views or reads. This puts the numbers into perspective and simplifies comparison between two periods, making it is easier to retrieve insights.

The **Topic Bubbles** artifact, described in section 4.4, shows what content is published, and what is consumed. It does so by showing the most published topics, next to the most extensively read topics. It has been attempted to map the topics to a fixed position, generic for every magazine. This did not work out well, due to the limitations that such a generic presentation would have, and the limited maturity of the topic matching at that

moment. For example, practically all topics can be considered part of Philosophy, and thus in any instance of this artifact, the center of topics moved toward Philosophy. A slightly modified, and simplified, artifact can still be useful to represent content, and compare content published with content consumed.

The **Bubble Up** artifact described, in section 4.5, shows how well the content that has been published, is read. It does so by positioning popular topics higher, while lowering less-popular topics, based on the UEI. It contains a 'worth publishing line' to enable publishers to identify topics that might not be worth publishing about. From a prototype it is learned, that it is important that topics are degraded when time passes. This way, topics extensively published about in the past, are not flooding today's overview. Currently, the x-axis represents time, although better options might exists. The Bubble Up artifact can probably best be used for a real time overview of the magazine, as it combines content with reading behavior.

The **Functional User Flows** artifact, described in section 4.6, shows the reader paths throughout the magazine. Only the three most visited paths are represented, along with a more-state and a leave-state. Expanding the states in the current visualization, can be overwhelming as soon as more than about two levels are expanded. In further development, this could be solved by automatically determining and expanding important paths, for example by integrating association rules. The Functional User Flows artifact can probably best be used for representing how readers flow to or from one specific article. In a modified form, combined with the Enriched View Trial, it can also be useful on a higher level.

Relating these findings and conclusions back to the literature, it is found that the ontology of Lengler and Eppler (2011) are not facilitating specifically designed artifacts well. The line between Information Visualization and Metaphor Visualization is not that clear, and only a limited number of visualizations are part of the ontology. This research basically proves the ontology could or should be extended, but how is open for discussion.

6.3 Customer Validation

In this section, we discuss and draw conclusions based on our third sub research question. This third sub question concerns the research pillar Customer Validation. First, let us reiterate the question:

SQ3: ***Why*** *and to which extent do these visualizations contribute to knowledge discovery by publishers in their magazines?*

6.3.1 Why this Question?

At the beginning of this research, it is stated that publishers need easy and intuitive access to digital publishing technology, so they can deliver the best content and improve reader engagement in order to stay competitive. In order to communicate the right information, it is important that not only the right information is delivered in the right way, but also that it is received as intended. Therefore, it is tested if the designed visualizations are able to communicate information is as intended.

6.3.2 How is it Answered?

In order to validate if the artifacts are able to communicate the right information, seven customer interviews were carried out in three phases. Next to these, internal information about the customers and the ability of the artifacts to transfer information was retrieved in Expert Sessions. The results are described in detail in chapter 1.

In the beginning of this research, 22 guiding questions have been formulated. These questions are rated in order to roughly identify the scope of the research. Some of the guiding questions are not answerable at all, because the data is not or cannot be available. Others are not answered because no relevant artifact made it

to the conceptual or functional phase. The research questions are assessed by what they mean in terms of feasibility (internal evaluation by creating the artifacts and executing internal discussions), the expert sessions and the customer evaluation. Based on these assessments, a rating from one (- -) to five (++) is assigned to get a rough feeling of the relevance of the artifact and how well the guiding questions are answered.

6.3.3 What are the Findings and Conclusions?

The five visualizations that are developed as artifacts, are shortly discussed below. The findings and conclusions here focus on the visualization and the findings while developing them. The Customer Evaluation is integrated in section 6.3.

For 12 out of the 22 guiding questions, one or more artifacts have been developed. For each artifact, the questions the artifact contributes to are listed, along with the other findings and conclusions. The complete list of questions with ratings can be found in Appendix C.

Generally, the wish for data greatly varies, and therefore also the extent to which the publishers are interested in the artifacts. Publishers that want to improve the information disclosure have a great whish for information, as have the publishers that want to sell advertisements. All traditional publishers are focused on profit, whereas the enterprise publishers are not; they are only interested in information disclosure. However, this distinction can blurry when (especially traditional) publishers get more aware what they learn from the information.

The **Swipe Patterns** artifact creates insight into how users exactly read an article. Publishers who are working on improving the content find it interesting to see precise behavior of users. Others are positive, but would likely not use it. Initially there was an internal wish to realize a generic heat map, as it was simpler, but as time passed, there turned out to be greater support for specific design. This artifact can be used to answer the guiding questions GQ14 and GQ20.

The **Enriched View Trail** artifact extends the currently used view trail and gives an overview of user behavior patterns. Customers expressed the wish to dive deeper, which leads us toward the functional User Flows Artifact. The current structure was sometimes hard to identify for customers. A dynamic layout, that represents the structure of the magazine, would likely work better. This artifact can be used to answer the guiding questions GQ15, GQ17, GQ20 and GQ21.

The **Topic Bubbles** artifact creates an overview of content in order to enable comparison between magazines, channels, etc. The experts are interested in the concept, but doubt the representation. All interviewed customers note that the visualization is too complex. This artifact can be used to answer the guiding questions GQ2, GQ3 and GQ4.

The **BubbleUp** artifact gives an overview of the content that is currently, real time, published and how successful it is. Although some customers found it more complex than needed, this artifact could create great insight for magazines that extensively publish. This artifact could contribute to answering the guiding question GQ3.

The **Functional User Flows** artifact visualizes how readers move through the magazine. The willingness of the customer to explore the complete paths differs. The use of icons that resemble the actual magazine is important. This artifact can be used to answer the guiding questions GQ11, GQ13, GQ17, GQ18, GQ19 and GQ21.

Relating this back to the literature, both edge cases exists for the tasks of visualizations by Shneiderman (1996). There are publishers who only wish to explore data at the highest level (“overview”), while others want

to have every detail and export it (“export”). Therefore, our results confirm that publishers also want to be facilitated on all the proposed levels of data facilitation.

On the other hand, some visualizations, especially the Topic Bubbles, were considered too complex, therefore traditional data visualizations, like bar and graph charts were mentioned as alternatives. It is probably a good idea to use traditional Data Visualizations where they are able to represent the information: at the lower levels of exploration.

6.4 Main Conclusions

In this section, the main conclusions are presented, along with some context toward this conclusion. First, let us restate the research question:

RQ: *How can we give publishers insight in their published content, the readers and their reading behavior in real time digital social magazines?*

To answer this question, we first state why the research was executed, then how it was executed and finally what the main conclusions are.

In the changing industry of digital publishing, imgZine creates a platform for real time social magazines. In the past years, both traditional and enterprise publishers have focused on competing and improving technology, instead of the creating, coordination and delivery of the right content for the right reader. Now the technology for exploring data moves toward an commodity, they need the right knowledge to do so.

In a Design Science study according to Peffers et al. (2007), we created five artifacts in order to explore the possibilities of improving and simplifying the access to content, readers and especially reading behavior. We did so by focusing on improving visualization. Based on customer interviews, expert sessions and development, we formulated the following three findings.

First, the **level of detail of information** publishers wish to have access to greatly differs among the publishers. Some publishers only want to know basic statistics, whereas others want to see reading behavior in great detail and on article level. Therefore, it is important that publishers are facilitated at all these levels, while still not becoming overwhelmed by the information they receive. Relating this back to the literature, both edge cases exist: the wish to explore data only at the highest level exists, while other want to dig into the lowest level of the tasks of visualization exploration as put forward by Shneiderman (1996). Therefore, our results confirm that publishers also want to be facilitated on all the proposed levels of data facilitation.

Second, **each level needs relevant visualizations** and these should be carefully selected in order to retain overview of the data on each level. The thee artifact related with reader path analysis (User Flows, Enriched View Trail and the Swipe Patterns) were well received, and give publishers insight into how readers behave throughout the magazines. However, when using visualizations to present detailed information on a higher level positive feedback declined. This is for example possible with the User Flows and the Swipe Patterns. Using these artifacts on a lower level was received more positively. The artifacts should be used on their own distinctive level of exploration, in order to retain overview.

Third, visualizations that **resemble the end-user product** clearly, were well understood among publishers. Artifacts that do not resemble any end-user product, were received negatively by publishers. Relating this back to the literature, it is expected that the boundary between Information and Metaphor Visualization (Lengler & Eppler, 2007) will be blurred, as the data gets more abstractions.

6.5 Limitations

In this section we look at the limitations and the possible threats to the validity of this research. According to Wieringa (2013), there are three types of validity that should be distinguished: construct, internal and external validity. For exploring possible threads to validity, these three types of validity are discussed below.

6.5.1 Construct Validity

Construct validity is concerned with the question of what we measure is a representation of the construct. In order to measure our construct, a great amount of different measures is used, as explained in detail in section

3.2. The main reason to do so, is to reach the best coverage of the different aspects (visualization and software development, data validation and customer validation). However, some possible threats are identified.

The customers, **publishers** are divided in both traditional publishers and enterprise customers. It is still believed that both are publishers, as the main role in the process, with regard to imgZine in specific, does not change. The publishers are likely more specialized at releasing a product to a specific audiences, where the enterprise customers are likely better at improving the experience to their customers. The validation of the functional artifact is only carried out to enterprise customers and although it is believed that the difference between the two groups is limited, it reduces the generalizability claim on traditional publishers.

The **maturity of development** of the artifacts greatly influences the customer understanding. As opposed to the artifacts related with reader path analysis, the Topic Bubbles and BubbleUp give insight into content. These artifacts are more positively received in the internal Expert Sessions, as they are in the external Customer Evaluations. These artifacts are probably not developed thoroughly enough for presentation towards the customers.

Privacy is outside of the scope of this research, but it can still influence the results. For example, when readers will tighten their requirements on privacy and what can be done with data about their behavior, they might influence (in)directly the publishers, and therefore the results and conclusions. In a side matter, most privacy problems relevant to this research, could probably be overcome by using k-anonymous algorithms to ensure individual readers is not traceable.

6.5.2 Internal Validity

Internal validity is concerned with the coherence between theory and empirical data. Our theories are concerned with the question that we can improve insight in data by delivering better visualizations.

First, on some relevant fields, only limited **literature review** has been carried out. Most notably, the pillar Customer Validation is not extensively supported by literature. This does include the fields like publishing industry and the tablets and app market. It is also noteworthy that the literature review on reading behavior and (visitor) flows could have been more extensive. Some fields of interest include Click Streams, Path Analysis, Association Rules, PetriNets and WF Formalization.

Second, when restarting the current research, another **research setup** would probably suit better. Especially one that allows comparison would improve the quality and generalizability of the results. This would make sense, as we also found in the literature that it is important for the understanding, that numbers and images are relative, and to interpret them in a usable way (McCandless, 2010). For example, such a setup would always compare a visualization from Google Analytics next to the proposed new visualization. This setup could be executed within the same methodology framework. However, the study is explicitly an explorative design study, which is a good practice in the field of Information Visualization. (Munzner, 2008)

Third, the **research approach** of Marchand and Peppard (2013) was followed. This could have been done more rigorously, by two means. First, in the initial phase of this research, more focus on building hypothesis and examine the data available, would have strengthened the research. Although this has been done, evaluating the research, more explicitly formulating the hypothesizes and conducting the Data Composition phase earlier, would have strengthened the research. Second, by intensifying the customer relations, and involve the customers in every step of the research. This could also be done by adding more people with knowledge on cognitive and behavioral science to the research, as is one of the required competences mentioned. However, this means the scope is narrowed to a more limited number of visualizations, possibly only one.

Last, some data available at imgZine is **inconsistent data** across magazines. The reason for this, is the dynamic evolvement of the imgZine publishing platform. During initial development, no specifications were made for

different characteristics that were logged, for example the events. Through the years, this has greatly improved, but only recently specifications were defined (see Appendix E). Some interpretations of the data in the functional prototype might not be correct for all magazines.

6.5.3 External Validity

External validity is concerned with the generalizability of the results and the conclusions.

By the nature of this research, we can only claim **limited generalizability**. Although we validated almost every step in the research, we do not have the quantity in Expert Sessions (n=4), Customer Interviews (n=7) and Artifact Development (n=1) to draw statistically valid conclusions. Also, the customers are not picked randomly. In a follow up research, it would be wise to test the designed artifacts in order to claim a greater generalizability. At best, suggested generalizability exists toward comparable solutions, especially with a comparable customer group, in the digital publishing industry.

6.6 Research Contributions

In this research, the **research approach** of Marchand and Peppard (2013) was followed. This research demonstrated an interesting case of an Big Data project. In such a project, practice and science moving toward each other, as developing becomes more and more a research process. At some points in the research, the focus shifted back to the traditional approach, which lead to a hurdle. This suggests, that the proposed approach is indeed a good approach. On the other hand, sometimes the tools need to be developed, before the next step can be executed. We thus propose a more parallel approach, where both the traditional and Big Data project approaches are integrated. In respect to this approach it is also important to note that the maturity of development of the artifacts greatly influences the customer understanding. Following the data-centric approach invites to rapid prototyping, which is a good thing on its own, but one should be careful not to work with too immature artifacts.

This research confirmed the **Tasks of Visualizations** of Shneiderman (1996) for the customers of imgZine. It suggests that this is also confirmed for the digital publishing industry. The wish for information by publishers span the edge cases of the tasks, and therefore all levels and thus all tasks of visualizations should exist.

By developing and evaluating five visualizations for examining reading behavior and content, an addition is made to the **ontology of visualizations** of Lengler and Eppler (2007). As for these Information Visualizations for the Publishing Industry, the use of metaphors is important, it is likely that the line between Information and Metaphor Visualization, as many by Lengler and Eppler, will blurry for Big Data Visualizations.

6.7 Future Research

In this section, possible future directions for research in academia are discussed. In the next sections, primarily future development and some research in practice are discussed.

The **integrating of BI&A 3.0 solutions** in the magazines (H. Chen et al., 2012). The integration of location-based services can greatly enhance the experience of the reader, by further personalization of the magazines. For example, if a reader is visiting a conference about a specific topic, the magazine could personalize by automatically recommending articles related to the topics of the conference. At the same time, facilitating location-based/BI&A 3.0, can deliver great insights into how readers use their products and article consumption can be placed into a greater, real world, context. At the same time, the need for more research and development of further visualizations evolves. Finally, as people have to agree with sharing their location in order to make use of location based services, it is important to have a clear goal and communicate it to the readers when enabling these technologies.

Content optimization and control is an interesting direction for future research. Currently, the recommendation engine proposed articles based on the behavior of readers. An publishers has only limited control over this process. It would be interesting to find out which factors a publishers *wants* to control, and which factors he or she has enough knowledge about to control them. This could be helpful for imgZine in specific, but also in general, as more and more websites and other content platforms get personalized.

For the in this study presented Topic Bubbles artifact, a **mapping of a network of topics on a 2D space** is needed. This was an unreachable goal within the time frame of this research. Finding a way to do so, might be an interesting direction for future research. That way, great networks of information could be presented in simple and easy accessible visualizations.

Information Visualization in Visual Analytics: Visual Analytics basically come in two forms: those tools which directly present data in a visual way, making it easier to interpret and those who already add an interpretation before the visual analysis is executed. We see a parallel with the distinction used in this work, between Data and Information Visualization. A future generation of Visual Analytics tools could well have integrated Information Visualizations, like the ones tested in this research. This can push the domain of Visual Analytics to the next level: with the integration of designed Information Visualization in Visual Analytics tools or Big Data Analytics, the accessibility for users with low or average technical knowledge level and the accurateness can be improved over automatic interpretation tools.

With the current Big Data solutions, the industry seems to move toward a more **research oriented** approach of developing. In order to do so, an exchange between research institutes and developing companies, is profitable. Both sides can profit from this, as the research institute can retrieve more new insights and empirical data from the companies, whereas the company has easier access to knowledge, can profit from the research network and can gain attention and significance by publishing. This probably means partnerships between research laboratories like ORTEC Living Data – of which imgZine is a part of – and universities or other knowledge institutes become more likely in the future.

6.8 Future Development

In this section, future development and some research in practice is discussed. As this is an explorative research, many directions and options for further development are possible. First, advised general development directions are discussed. Second, the future development of the individual artifacts is discussed.

Smart Dashboard: after analyzing the magazine usage by using the Analytics Dashboard, publishers should be able to stay in control over their magazines, even when the technology is abstract and complex. In order to facilitate this, we propose that decisions are limited to simple choices between two or three directions. Smart algorithms will automatically group readers and content. Recommendations about the implications of the decisions are also considered interesting to integrate. This can be seen as a Decision Support System (DSS) for the Big Data purposes.

Artifact Development: of the five artifacts in this research, four artifacts are useful and feasible to develop in the short term for the Analytics Dashboard of imgZine. These four artifacts are the Enriched View Trail, Functional User Flows, Swipe Patterns and Bubble Up. The development of the artifacts is discussed:

> First, integration of the reading behavior artifacts is proposed. The three artifacts concerned with reading behavior could best be used on one or more levels of the tasks of visualizations (Shneiderman, 1996) to deliver the best integrated user experience. This means further developing the Functional User Flows for presenting information on low-level reader flows toward and away from specific articles. On this same level Swipe Patterns could help finding interesting insights. The Enriched View Trail could be very helpful on the highest level of information exploration. The

comparison element is an important feature to use; and this could be a good starting point for development.

Second, the Functional User Flows artifact is useful on two levels: both for facilitating expansion of the View Trail, as well as on the level of the reader paths toward and away from specific articles. The Functional User Flows artifact does not include all intended designed characteristics. First, the intended design does expand the first four levels of the most followed path directly when opening the visualization. It would also align these on the top in one row, making it more similar to the Enriched View Trail. Second, it would include a more expressive visualization of the percentage of readers that follows a path. An example of such a visualizations could be a paper stack metaphor for each state. It is expected that designing an artifact that would include these characteristics would increase the usability.

Third, the Bubble Up artifacts could be a distinctive view of the dashboard, which could especially be interesting to show the current content published and consumed real time, for example on an in-company electronic billboard. It could also be useful as explorative tool on a certain point in time. For both of these solutions the data on the x-axis should be reconsidered, as better solutions might well exist.

Last, The Topic Bubbles is in the currently presented format not useful, as the concept of fixed position topics is too complex/abstract to build. In simplified form, it can be useful. Especially the idea to represent the topics in a way that they can be laid over each other can lead to understanding and insight.

Diversification: imgZine is part of the ORTEC Living Data lab, a lab with many start-ups involved in creating solutions based on data analysis. There are many options for diversification. We give three examples:

First, a soccer player analysis solution is currently in development, for which certain visualizations can be suitable. For example, the Swipe Patterns could be adjusted in order to represent the trails of soccer players. Or, the User Flows artifact could be used to represent a trail of particular moves of players, in order to identify successful and less successful combinations. Furthermore, one could even think of using the content analysis artifacts for representing the important players in the game.

Second, another start-up is involved in advertisement and lead analysis, based on data patterns. The current dashboard greatly relies on a high knowledge level of the person who is controlling it. However, the advertisement and marketing market is also moving toward commodity and so will their customers. An easy and intuitive way of knowledge discovery enables these new customers to get the best out of the technology.

Third, the integration of BI&A 3.0, as proposed in section 6.7, could lead to a combination of the services delivered by a location-based services firm at ORTEC Living Data and the magazines from imgZine. This way, the artifacts designed could also be used to analyze how users are using location-based services.

6.9 Personal Reflection

Each work deserves a reflection. This one is mine. You have made it till the end of my Master's thesis. Or, you just skipped everything, just to read my opinion. And I love to give opinions, so this should work out well.

I enjoyed having the freedom to set up my own research, along with great people, and the great supervisors that I was able to choose myself. It was a great learning experience and I do not regret my choices. I especially enjoyed learning about the aspects of visualizations, like color coding and design patterns, and designing them myself. I also enjoyed creating prototypes in JavaScript more than I had expected.

What I enjoyed less were the customer interviews. Although I had expected otherwise, working with qualitative data is not what gets me out of bed in the morning.

Anyway, a thesis is a written document. And that is something I didn't enjoy at all. So, no matter how much I learned, I would have preferred to use another reporting method. Since websites like TED(x) are becoming increasingly popular, and people spend less time reading long articles*, I think the future lies in video presentations. I propose no more written theses are made, but full hour college presentations.

Bernard van der Wees

December 2013

** That would be a great hypothesis to test using the new Analytics Dashboard.*


imgZine

References

Adformatie. (2013, November 2). Oplage Sanoma daalt harder dan de markt. Retrieved November 2, 2013, from http://www.adformatie.nl/nieuws/bericht/oplage-sanoma-daalt-weer-harder-dan-de-markt/

Agrawal, D., Das, S., & El Abbadi, A. (2011). Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology* (pp. 530–533). Retrieved from http://dl.acm.org/citation.cfm?id=1951432

Apache Software Foundation. (2012). Welcome to Apache™ Hadoop®! Retrieved November 19, 2013, from http://hadoop.apache.org/index.html

Apache Software Foundation. (2013). Hadoop Releases. Retrieved November 19, 2013, from http://hadoop.apache.org/releases.html

Appcelerator Inc. (2013). Titanium Mobile Application Development. Retrieved August 22, 2013, from http://www.appcelerator.com/platform/titanium-platform/

Behrens, C. (2008a). InfoDesignPatterns.com. Retrieved May 17, 2013, from http://www.niceone.org/infodesignpatterns/index.php5#/home.php5

Behrens, C. (2008b, February). *The Form of Facts and Figures* (Master of Arts in Design in the Interface Design program). Potsdam University of Applied Sciences, Postdam.

Bioinformatics Laboratory. (2013). Orange – Data Mining Fruitful & Fun (Version 2.7). Ljubljana, Slovenia: Faculty of Computer and Information Science, University of Ljubljana. Retrieved from http://orange.biolab.si/

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2012). The meaningful use of big data: four perspectives–four challenges. *ACM SIGMOD Record*, *40*(4), 56–60. Retrieved from http://dl.acm.org/citation.cfm?id=2094129

Bloem, J., Van Doorn, M., Duivestein, S., & van Ommeren, E. (2012). Creating clarity with Big Data. *Sogeti VINT*. Retrieved from http://www.sogeti.se/upload/SV/Kalendarium/Dokument/Big-data1.pdf

Bremer, K. (2007, May 11). Is a picture worth a thousand words? Improving decision making with visual analytics. Master's Thesis. Retrieved May 13, 2013, from http://eprints.eemcs.utwente.nl/9847/

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88–98.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Chen, Y., Alspaugh, S., & Katz, R. (2012). Interactive analytical processing in big data systems: a cross-industry study of MapReduce workloads. *Proceedings of the VLDB Endowment*, *5*(12), 1802–1813. Retrieved from http://dl.acm.org/citation.cfm?id=2367519

Clements, P., & Bass, L. (2010). Using Business Goals to Inform a Software Architecture. In *Requirements Engineering Conference (RE), 2010 18th IEEE International* (pp. 69–78). doi:10.1109/RE.2010.18

De Morgen. (2013, November 6). Mediahuis schrapt 205 banen. Retrieved December 9, 2013, from http://www.demorgen.be/dm/nl/16340/Media/article/detail/1735834/2013/11/06/Mediahuis-schrapt-205-banen.dhtml

Dean, J., & Ghemawat, S. (2004). MapReduce: simplified data processing on large clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*. Retrieved from http://research.google.com/archive/mapreduce.html

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of*

the ACM, *51*(1), 107–113.

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, *54*(4), 585–612. doi:10.1006/ijhc.2001.0458

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases. *AI Magazine*, *13*(3). Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1011

Google Ideas. (2013). Google Ideas Projects. Retrieved October 30, 2013, from http://www.google.com/ideas/projects/

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *Management Information Systems Quarterly*, *37*(2), 337–355.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A Tour Through the Visualization Zoo. Retrieved May 17, 2013, from http://hci.stanford.edu/jheer/files/zoo/

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Q.*, *28*(1), 75–105.

Hibbard, B. (2004). The top five problems that motivated my work [data visualisation]. *Computer Graphics and Applications, IEEE*, *24*(6), 9–13. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1355886

Iliinsky, N. (2010). On Beauty. In *Beautiful Visualization* (pp. 1–14). O'Reilly Media.

imgZine. (2013a). imgZine – platform for publishing real time social magazines | About. Retrieved June 24, 2013, from http://imgzine.com/about/

imgZine. (2013b, September). Events Specification and Wishlist (internal documentation).

Johnson, C. (2004). Top scientific visualization research problems. *Computer graphics and applications, IEEE*, *24*(4), 13–17. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1310205

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1–8. doi:10.1109/2945.981847

Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in Visual Data Analysis. In *Tenth International Conference on Information Visualization, 2006. IV 2006* (pp. 9–16). doi:10.1109/IV.2006.31

Koblin, A., & Klump, V. (2010). Flight Patterns: A Deep Dive. In *Beautiful Visualization* (pp. 91–102). O'Reilly Media.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, *5*(12), 2032–2033. Retrieved from http://dl.acm.org/citation.cfm?id=2367572

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety.

Laurila, J. K., Gatica-Perez, D., Aad, I., Blom, J., Bornet, O., Do, T.-M.-T., … Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. In *Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing* (pp. 1–8). Retrieved from http://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf

Lee, E. (2012, May 23). A Taxonomy of Data Visualization. *Visualizing.org*. Retrieved May 17, 2013, from http://visualizing.org/stories/taxonomy-data-visualization

Lengler, R., & Eppler, M. J. (2007). Towards a periodic table of visualization methods for management. In

IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA.

Lengler, R., & Eppler, M. J. (2011, November 9). A Periodic Table of Visualization Methods. Retrieved May 17, 2013, from http://www.visual-literacy.org/periodic_table/periodic_table.html

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Information*. Princeton Architectural Press.

Marchand, D. A., & Peppard, J. (2013). Why IT Fumbles Analytics. *Harvard Business Review*, *91*(1), 104–112.

McCandless, D. (2009). *Information is beautiful*. London: Collins.

McCandless, D. (2010). *The beauty of data visualization*. TEDGlobal 2010. Retrieved from http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

McCandless, D. (2013). Information Is Beautiful | Visualizations. Retrieved May 3, 2013, from http://www.informationisbeautiful.net/visualizations/

Meertens, L. O. (2013). *From Business Modelling to Enterprise Architecture* (PhD Dissertation). University of Twente, Enschede. Retrieved from http://www.lmeertens.nl/thesis/phdthesisv57.pdf

Munzner, T. (2008). Process and Pitfalls in Writing Information Visualization Research Papers. In *Information Visualization: Human-Centered Issues and Perspectives* (Vol. 4950, pp. 134–153). Springer. Retrieved from http://www.cs.ubc.ca/labs/imager/tr/2008/pitfalls/

Nielsen, J. (1997, October 1). How Users Read on the Web. Retrieved April 23, 2013, from http://www.nngroup.com/articles/how-users-read-on-the-web/

Nielsen, J. (2006, April 17). F-Shaped Pattern For Reading Web Content. Retrieved April 23, 2013, from http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

Nielsen, J. (2010, March 22). Scrolling and Attention. Retrieved April 23, 2013, from http://www.nngroup.com/articles/scrolling-and-attention/

Novet, J. (2013, June 6). Facebook unveils Presto engine for querying 250 PB data warehouse. *Gigaom*. Retrieved June 7, 2013, from http://gigaom.com/2013/06/06/facebook-unveils-presto-engine-for-querying-250-pb-data-warehouse/

Nunamaker Jr, J. F., Briggs, R. O., Mittleman, D. D., Vogel, D. R., & Balthazard, P. A. (1996). Lessons from a dozen years of group support systems research: a discussion of lab and field findings. *Journal of management information systems*, 163–207.

ORTEC. (2013). Business Units. Retrieved December 4, 2013, from http://www.ortec.nl/business-units.aspx

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45–77.

Pfauth, E. J. (2013, November 1). De eerste maand van De Correspondent in cijfers. Retrieved November 2, 2013, from https://decorrespondent.nl/282/de-eerste-maand-van-de-correspondent-in-cijfers/12811027350-2d249a45

Reijerman, D. (2013, August 26). Uitgever PCM, Computer Idee, PC-Active en PU vraagt faillissement aan. Retrieved December 9, 2013, from http://tweakers.net/nieuws/90913/uitgever-pcm-computer-idee-pc-active-en-pu-vraagt-faillissement-aan.html

Sanoma. (2013, October 31). Sanoma to rationalise Dutch magazine portfolio as part of consumer media redesign. Retrieved November 2, 2013, from http://www.sanoma.com/en/news/sanoma-rationalise-dutch-magazine-portfolio-part-consumer-media-redesign

SAS Institute Inc. (2013). Big Data – What Is It? Retrieved June 7, 2013, from http://www.sas.com/big-data/

Shankland, S. (2008, May 30). Google spotlights data center inner workings. *CNET News*. Retrieved November 19, 2013, from http://news.cnet.com/8301-10784_3-9955184-7.html

Shapiro, M. (2010). Once Upon a Stacked Time Series. In *Beautiful Visualization* (pp. 1–14). O'Reilly Media.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *, IEEE Symposium on Visual Languages, 1996. Proceedings* (pp. 336–343). doi:10.1109/VL.1996.545307

Snijders, C., Matzat, U., & Reips, U.-D. (2012). "Big Data": Big Gaps of Knowledge in the Field of Internet Science (Editorial). *International Journal of Internet Science*, *1*(7), 1–5.

Staughton, K. (2012). *Tablet magazines and the affects on the magazine industry*. California Polytechnic State University.

Steele, J., & Iliinsky, N. (2010). *Beautiful Visualization*. O'Reilly Media.

Thomas, J. J., & Cook, K. A. (Eds.). (2005). Illuminating the Path. PNNL: Information Visualization and Visual Analytics. Retrieved from http://vis.pnnl.gov/

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., … Murthy, R. (2010). Hive - a petabyte scale data warehouse using Hadoop. In *2010 IEEE 26th International Conference on Data Engineering (ICDE)* (pp. 996–1005). doi:10.1109/ICDE.2010.5447738

TNO. (2013). Dynamiek op de Nederlandse arbeidsmarkt 2013. Retrieved December 9, 2013, from http://www.monitorarbeid.tno.nl/publicaties/dynamiek-op-de-nederlandse-arbeidsmarkt-2013

Van der Wees, B. J., & Moonen, H. (2011). The Potential of In-train Crowdsourcing. In *eFuture: Creating Solutions for the Individual, Organisations and Society* (Vol. Research Volume). Bled, Slovenia.

Visual.ly. (2013). Infographics & Data Visualization | Visual.ly. Retrieved May 3, 2013, from http://visual.ly/

VPRO. (2013, oktober). Uw persoonlijke data zijn goud waard. *Tegenlicht*. NPO. Retrieved from http://www.uitzendinggemist.nl/afleveringen/1375980

Wiederkehr, B., Siegrist, C., Stucki, J., Gassner, P., & Schmid, C. (2013). Datavisualization.ch Selected Tools. Retrieved May 3, 2013, from http://selection.datavisualization.ch/

Wieringa, R. (2009). Design science as nested problem solving. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (pp. 8:1–8:12). New York, NY, USA: ACM. doi:10.1145/1555619.1555630

Wieringa, R. J. (2013, February). *Slides from Design Science methodology (course)*.

Wong, P. C., & Thomas, J. (2004). Guest Editors' Introduction--Visual Analytics. *IEEE Computer Graphics and Applications, 24(5):20-21*, *24*(5). doi:10.1109/MCG.2004.39

WPG Uitgevers. (2013, November). WPG Uitgevers kondigt reorganisatie aan. Retrieved November 19, 2013, from http://www.wpg.nl/nieuws/nieuwsbericht/WPG-Uitgevers-kondigt-reorganisatie-aan.htm

Yin, R. K. (2008). *Case Study Research: Design and Methods* (4th ed.). SAGE Publications, Inc.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., … Keim, D. (2012). Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (pp. 173–182). doi:10.1109/VAST.2012.6400554

Zikopoulos, P., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2012). Understanding big data. *New York et al: McGraw-Hill*.

Appendix A List of Abbreviation and Terms

Analytics Dashboard	The part of the publishers dashboard where information and statistics about usage can be found. The analytics dashboard and it's users are the problem context of this research.
BI&A	"Business Intelligence & Analytics" (H. Chen, Chiang, & Storey, 2012)
BYOD	Bring Your Own Device.
Configuration Dashboard	The part of the publishers dashboard used to configure the magazines; e.g. sources for the articles, the channels containing them and so on, can be configured here.
Content	*(media)* ~: Articles, including images, video's and other media and the channels in which they are grouped, which are (planned) to be published using the magazines.
Customers	(Corporate) ~: See *Publishers*. Two types of publishers are distinguished: *traditional publishers* and *enterprise publishers*.
Data Visualization	"visual representations of quantitative data in a schematic form" (Lengler & Eppler, 2007)
Enterprise publishers	Customer group of imgZine, who create magazines for their own employees, or for their customers, most often for information delivery. See also *traditional publishers*.
ETL	Extract, Transform, Load
Information Visualization	"the use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by the user as they proceed working with it" (Lengler & Eppler, 2007)
Knowledge Discovery	"nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley, Piatetsky-Shapiro, & Matheus, 1992)
Publishers	traditional publishers or enterprises who bought their own magazine at imgZine. These are the primary users of the dashboard, configuring their magazine in the configuration dashboard and retrieving information and insight in the analytics dashboard. Two types of publishers are distinguished: *traditional publishers* and *enterprise publishers*.
Readers	The individuals who read the magazines and are the end-users of the delivered magazines.
Reading Behavior	All data, actions and other characteristics that are related to what the reader *does* when reading the magazines. Examples are the sequences of articles read and the time spend on reading.

Real time social magazine	A (digital) app, often multi-platform, which loads articles from sources, as soon as they are released (in case of push), or with a small delay (in case of pull).
Recommendations	Advice for improving content or the magazines, based on the information that is retrieved through data extraction and visualization.
Smart Intranet	An intranet that combines collaboration, document management, social and decision making functions in an easy way. (Rhoads, 2013)
SQL	See *Structured Query Language*.
Structured Query Language	A special-purpose programming language for managing and retrieving data from databases.
Topic Extraction	The automatic matching process between articles and topics, based on the text that is related to them, using text analysis technology.
Traditional publishers	Customer group of imgZine, who create magazines for external readers. Examples are magazines and newspapers. See also *enterprise publishers*.
UEI	See *User Engagement Index*.
User Engagement Index	A combination of metrics used to determine the actual engagement a user has with a certain article or a set of articles. The basic metric is calculated by correcting the reading speed for the article length. It is used both internally, but also as relative value to the customer for distinguishing the amount of attention. It does not say anything about if the user actually likes the article or not.
White-label magazines	Magazines which are branded according to the wishes and styles of the customers.

Appendix B List of Stakeholders and Goals

In this appendix, all stakeholders for this research and their goals are identified using the stakeholder types proposed by Clements and Bass (2010). This list also includes stakeholders that are on the boundaries of relevant.

Table B.1: Complete list of stakeholders and their goals

Type	Stakeholder	Goal
Customers	(Corporate) Customers	• Retrieve clear and advanced usage statistics; • Understand if content is popular; • Adept content to customer needs.
Customers	Publishers	• Find ways to optimize revenue generation
Customers	Enterprises	• Inform our users or employees; • Improve customer relations; • Generate revenue indirectly.
Employees	CEO	• Improve insight in data by improved visualizations; • Patent possible new (visualization) technologies; • Give content providers insight in what content is popular and what is content is missing
Employees	COO	• Keep the software architecture workable; • Do not threat operations; • Improving commercial position; • Improve recommendation and information statistics; • Generate and sustain competitive advantage
Employees	Project Manager	• Revealing missed changes by customers; • Enable cross- and up-selling changes in the future
Employees	Developers	• Design challenging technologies; • Do not interrupt regular operations; • Freedom to develop wished skills
Governments	Privacy Watchdog	• Secure the personal data of individual citizens
Investors	imgZine	• Improve opportunities for new products to be cross-sold.
Political groups	*none*	
Suppliers	Apple	• Have influence on the usage and sells of the AppStore
Suppliers	Microsoft	• Have influence on the usage and sells of the Windows (Phone) Marketplace
Suppliers	Google	• Have influence on the usage and sells of the Play Store
Trade Associations	Stimuleringsfonds voor de Pers	• Improve quality, diversity and independency by means of innovation in the journalism/publishing industry
Communities	*none*	

Appendix C List of Guiding Questions

Table C.2: Complete list of guiding questions

Nr.	Guiding Question	Category	Attractiveness	Relevant Artifact*	Feasibility	Expert Sessions	Customer Evaluation
GQ1	Is there a difference between magazines by younger and older companies?	Content	--	n/a	+/-	n/a	--
GQ2	Which sources are better read than others?	Content	+	Topic Bubbles	++	++	+/-
GQ3	Are some articles better not published at all?	Content	++	Topic Bubbles, BubbleUp	+	++	+
GQ4	Do certain sources even degrade reader engagement for a complete magazine?	Content	++	Topic Bubbles	-	+	+
GQ5	Could an optimal magazine automatically be composed out of available sources?	Content	+	n/a	+/-	++	n/a
GQ6	Could sources be selected based on usage pattern?	Content, Reading Behavior	++	n/a	+	++	n/a
GQ7	Can we determine which articles people want to read, but do not read?	Content, Reading Behavior	++	n/a	-	n/a	n/a
GQ8	Can we determine the age of readers?	Readers	+/-	n/a	-	--	n/a
GQ9	Which type of readers do we have? / Can we distinguish different groups of readers?	Readers, Reading Behavior	+	User Groups*	+	-	+/-
GQ10	What are the characteristics of these reader groups? / Can we distinguish characteristics like the educational level, income and gender of readers?	Readers, Reading Behavior	+	User Groups*	-	-	+
GQ11	Can we distinguish readers based solely on their behavior patterns through the magazines?	Readers, Reading Behavior	++	User Flows	+/-	++	++
GQ12	At which days and times are the magazines most intensely read?	Reading Behavior	++	n/a**	++	++	++
GQ13	At which locations are the magazines read the most?	Reading Behavior	+/-	User Flows	+/-	+/-	+/-
GQ14	Do readers make different decisions when reading in landscape or in portrait orientation?	Reading Behavior	+	Swipe Patterns	+	-	+
GQ15	Why do people drop off directly? And how many of them do so?	Reading Behavior	+/-	Enriched View Trail, User Flows	-	-	+/-
GQ16	Why do people download an app, and never open it at all?	Reading Behavior	+	n/a	--	-	+/-
GQ17	How often do people open magazines without reading anything?	Reading Behavior	+	User Flows, Enriched View Trail	++	+/-	+
GQ18	Do people read articles in a	Reading	++	User Flows	+	-	++

	specific order?	Behavior					
GQ19	Do people choose for specific articles to read, or do they pick them (relatively) random?	Reading Behavior	++	User Flows	+/-	+	+/-
GQ20	Can we distinguish different reading behavior between different devices/device types?	Reading Behavior	+	Swipe Patterns, User Flows, Enriched View Trail	+	+/-	++
GQ21	Can we distinguish working hour patterns? Are these magazine specific?	Reading Behavior	+	User Flows, Enriched View Trail	++	++	+
GQ22	How and why do people read articles less or more extensive?	Reading Behavior	++	n/a	-	+	++

* *Some artifacts are discontinued before the reaching the conceptual phase.*

** *Some questions can (almost) be answered using the old Analytics Dashboard.*

n/a *not available/no artifact.*

Appendix D Overview of Research Process

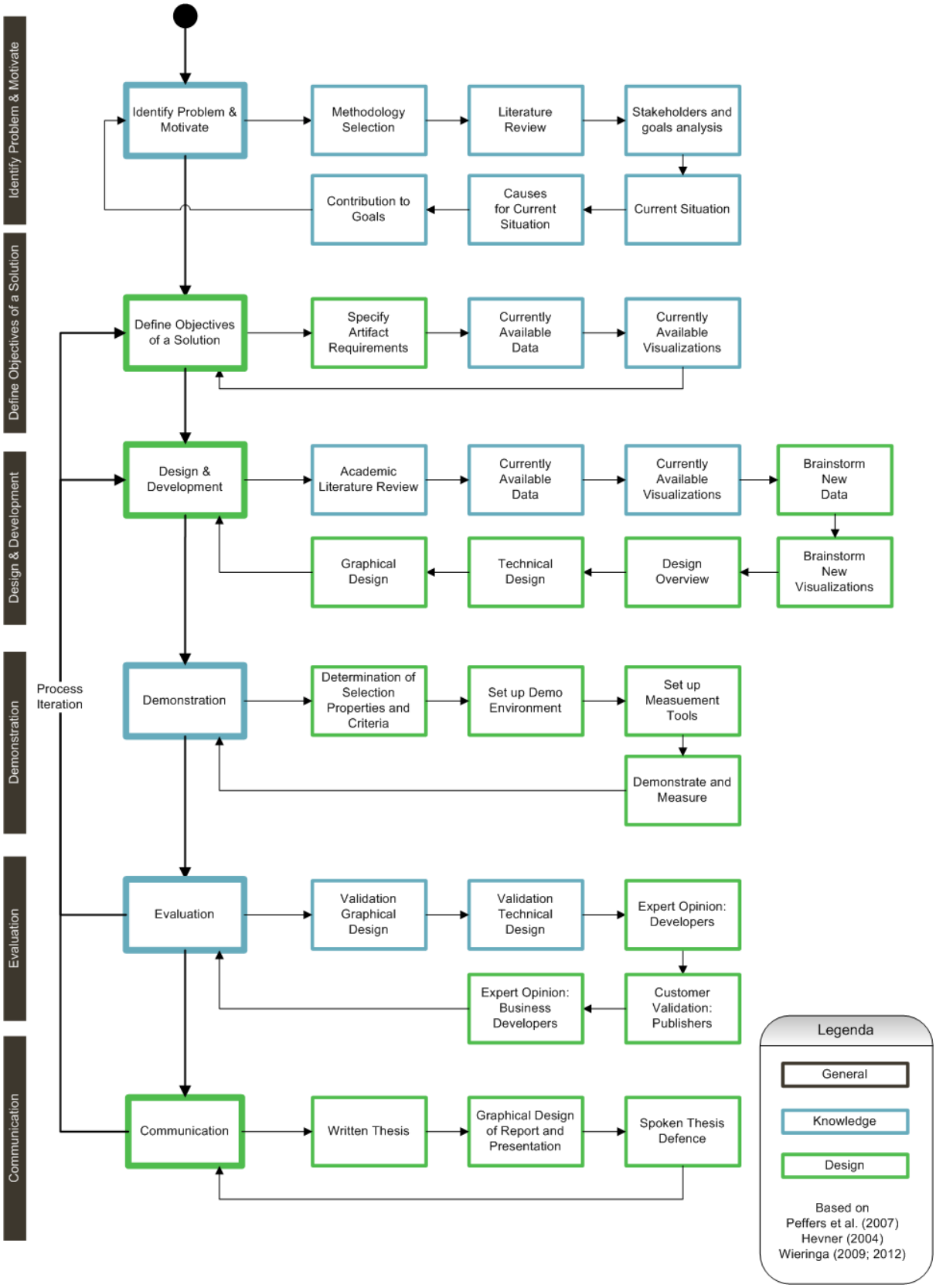


Figure D.1: Extensive research design

Appendix E Event Specifications

This appendix contains the improved event specification. This specification did not yet exist when this research started. This appendix is literally cited from internal documentation. (imgZine, 2013b)

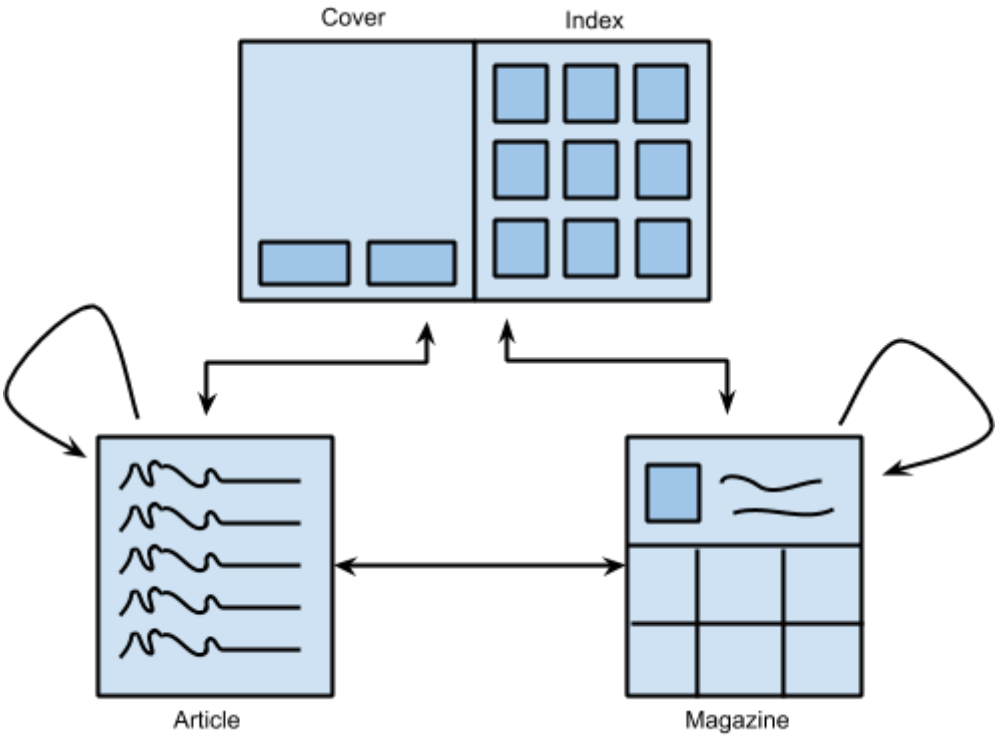


Figure E.2: Overview of possible transitions between page types

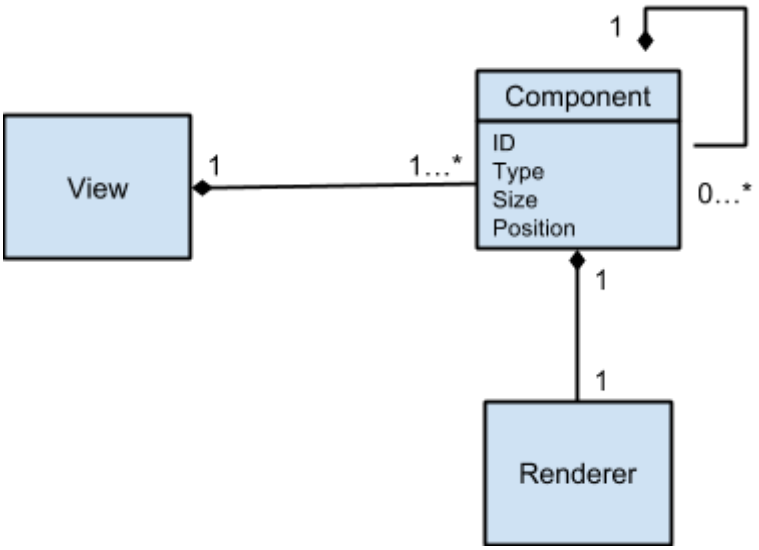


Figure E.3: Overview of database structure

A message sent from an application to *inbox.py* consists of the following data (Table E.3).

Table E.3: Event message data

NAME	DESCRIPTION
magazine	
iz	API key
os_name	
os_version	
device_id	
device_model	
app_version	
req	

Table E.4: Event parameters

NAME	DESCRIPTION	VALUES
msg_type		One of: *event, rating, comment, feedback, invite*
udid	Unique device-application ID.	
cdate	Event creation time stamp.	
magazine	Unique magazine ID.	
ip_address	IP-address of the current connection of the client device.	
msg		
referer	Component that triggered this event. Same set of values as renderer.	
type	Action that triggered the event.	view, favorite_add, app_start, etc
item	Number associated with the current type.	Article ID; [...]
item_type		platform_article; platform_channel; user_channel; advertorial; etc
renderer	Layout for the current item.	Cover; Index; BoxMagazine; ListMagazine; DetailMagazine; LegoMagazine; VerticalMagazine; and others
children	A list of (*id*, *type*, *renderer*, *box/index*, *children*)	
extra		

Note on naming conventions: renderer values should be written in upper camel case notation; everything else, in lower case with words separated by underscores.

Table E.5: Event types

view		
app_start		
app_close		

Table E.6: Event specific considerations

Trigger	Application started or resumed
type	app_start
item	-
item type	-
renderer	-
msg	-
extra	*device*; *resolution; size; screenWidth (px); screenHeight (px); os; model (e.g.: HTC Sensation Z710e)*
Trigger	Application closed or paused
type	app_close
item	-
item type	-
renderer	-
msg	Duration of the visit (time, in milliseconds, from app_start to app_close)
extra	
Trigger	Cover page view
type	view
item	-
item type	-
renderer	Cover
msg	-
children	Note: ideally, the cover should also have a list of children.
extra	-
Trigger	Index page view
type	view
item	-
item type	-
renderer	Index
msg	-
children	Note: ideally, the index should also have a list of children.
extra	-
Trigger	Magazine view
type	view
item	(Channel's / Collection ID)
item type	e.g. platform_channel, user_channel, fav_folder
renderer	*Magazine
msg	Duration of the view in milliseconds.
children	A list of (*id*, *type*, *renderer*, *box/index*, *children*)
extra	*index* (i.e. page number); *orientation* (*portrait* or *landscape*)
Trigger	Article detail view
type	view
item	(Item's Id)
item type	e.g. platform_channel
renderer	*Magazine
msg	Duration of the view in milliseconds.
children	A list of (*id*, *type*, *renderer*, *box/index*, *children*)
extra	*collection* (ID of the parent collection, e.g. channel); *collection_type; class* (one of: *read*, *swipe* or *bug*)