DOI: 10.1002/pra2.392

POSTERS

Utilizing HTML-analysis and computer vision on a corpus of website screenshots to investigate design developments on the web

Thomas Schmidt | Anastasiia Mosiienko | Raffaela Faber | Juliane Herzog | Christian Wolff

Media Informatics Group, University of Regensburg, Regensburg, Germany

Correspondence
Thomas Schmidt, Media Informatics Group, University of Regensburg, Regensburg, Germany.
Email: thomas.schmidt@ur.de

Abstract
We present preliminary results of a project investigating the design development of popular websites between 1996 and 2020 via HTML analysis and basic computer vision methods. We acquired a corpus of website screenshots of the current top 47 popular websites. We crawled a snapshot of every month of these websites via the wayback machine of the Internet Archive platform since the time snapshots are stored to gather 7,953 screenshots and HTML pages. We report upon quantitative analysis results concerning HTML elements, color distributions and visual complexity throughout the years.

KEYWORDS
colors, html, visual complexity, web design, web history, websites

1 | INTRODUCTION

The World Wide Web has become an important part of modern media infrastructure and society. The web itself has also become an object of research in human-computer-interaction (Jørgensen & Myers, 2008) but also cultural and media studies (Brügger, 2012). Investigating the history of web interfaces is an important task for web and media historians but can also give current web designers inspiration on how developments might continue. One important aspect of this research area is the archival of the web and the preservation of this digital heritage has been addressed by the UNESCO.[1] This led to the platform Internet Archive,[2] which, via its Wayback Machine[3] intends to archive the web and enables design researchers to investigate trends on large-scale corpora. We present work-in-progress results of a project investigating web design developments via quantitative and qualitative analysis. We report on our current approach and the first quantitative results we acquired.

2 | METHODS

2.1 | Corpus-acquisition

We decided to analyze the top 50 most popular websites as of December 2019 according to the analytical platform *Alexa*.[4] We filtered out any adult websites, which resulted in a list of 47 websites. We acquired one snapshot per day of these websites stored in the Wayback Machine, if a snapshot was available for the timespan from 1996 to 2020. A snapshot is a stored representation of the website for a given time. This results in a corpus of 151,682 snapshots. However, all websites are represented rather unequally with the most popular and oldest being most

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

TABLE 1 Websites of the corpus and number of snapshots

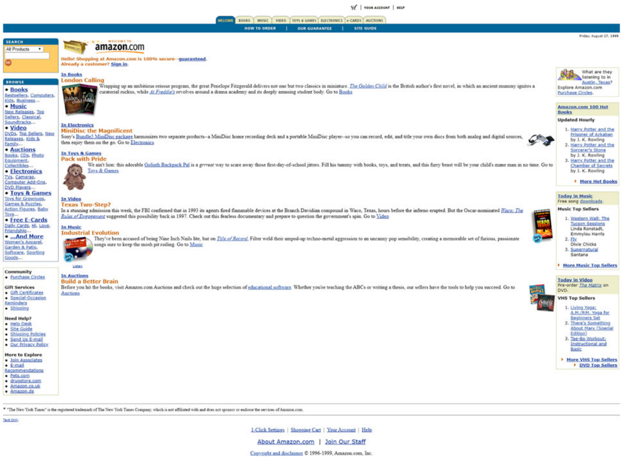
Website	#snapshots	Website	#snapshots	Website	#snapshots	Website	#snapshot
360.cn	152	Facebook.com	200	Office.com	98	Twitter.com	183
Aliexpress.com	121	Google.co.in	194	Okezone.com	155	Vk.com	169
Alipay.com	121	Google.com.hk	199	qq.com	202	Weibo.com	103
Amazon.co.jp	161	Google.com	242	Reddit.com	179	Wikipedia.org	191
Amazon.com	203	Instagram.com	114	Sina.com	218	Wordpress.com	184
Apple.com	242	Jd.com	132	Sohu.com	227	Xinhuanet.com	228
Babytree.com	184	Live.com	102	Stackoverflow.com	138	Yahoo.co.jp	245
Baidu.com	216	Login.tmall.com	92	Taobao.com	126	Yahoo.com	255
Bing.com	159	Microsoftonline.com	124	Tianya.cn	169	Yandex.ru	220
Blogspot.com	107	msn.com	241	Tmall.com	128	Youtube.com	178
Csdn.net	214	Naver.com	229	Tribunnews.com	110		
Ebay.com	240	Netflix.com	153	Twitch.tv	105		

TABLE 2 Website distribution per year

Year	#snapshots	Year	#snapshots
1996	10	2009	378
1997	23	2010	412
1998	26	2011	487
1999	70	2012	494
2000	150	2013	500
2001	183	2014	487
2002	182	2015	491
2003	284	2016	484
2004	317	2017	503
2005	349	2018	498
2006	362	2019	501
2007	394	2020	30
2008	338		

frequent. For this study, we limited this corpus to one snapshot per month per website (the first available snapshot per month) to avoid problems with the unequal distribution. This subcorpus consists of 7,953 snapshots (Tables 1 and 2) of which we scraped the HTML and took a screenshot as TIFF-File with a width of 1920 pixels and height according to the size of website. To enable comparisons, we sliced the screenshots at 3000 pixels height.

The sample size for the early years is rather limited. However, beginning 2003 the sample size is more representative with around 300 snapshots per year. Figures 1 and 2 show two snapshots of amazon.com.



Friday, August 27, 1999

Hello! Shopping at Amazon.com is 100% secure--guaranteed.
Already a customer? Sign in.

In Books
London Calling
Wrapping up an ambitious reissue program, the great Penelope Fitzgerald delivers not one but two classics in miniature. *The Golden Child* is the British author's first novel, in which an ancient mummy ignites a curatorial ruckus, while *At Freddie's* revolves around a drama academy and its deeply amusing student body. Go to Books

In Electronics
MiniDisc the Magnificent
Sony's Bundle5 MiniDisc package harmonizes two separate products--a MiniDisc home recording deck and a portable MiniDisc player--so you can record, edit, and title your own discs from both analog and digital sources, then enjoy them on the go. Go to Electronics

In Toys & Games
Pack with Pride
We ain't lion: this adorable Goliath Backpack Pal is a *grrreat* way to scare away those first-day-of-school jitters. Fill his tummy with books, toys, and treats, and this furry beast will be your child's mane man in no time. Go to Toys & Games

In Video
Texas Two-Step?
In a stunning admission this week, the FBI confirmed that in 1993 its agents fired flammable devices at the Branch Davidian compound in Waco, Texas, hours before the inferno erupted. But the Oscar-nominated *Waco: The Rules of Engagement* suggested this possibility back in 1997. Check out this fearless documentary and prepare to question the government's spin. Go to Video

In Music
Industrial Evolution
They've been accused of being Nine Inch Nails lite, but on *Title of Record*, Filter weld their amped-up techno-metal aggression to an uncanny pop sensibility, creating a memorable set of furious, passionate songs sure to keep the mosh pit roiling. Go to Music

In Auctions
Build a Better Brain
Before you hit the books, visit Amazon.com Auctions and check out the huge selection of educational software. Whether you're teaching the ABCs or writing a thesis, our sellers have the tools to help you succeed. Go to Auctions

- Books
Bestsellers, Computers, Kids, Business...
- Music
New Releases, Top Sellers, Classical, Soundtracks...
- Video
DVDs, Top Sellers, New Releases, Kids & Family...
- Auctions
Books, CDs, Photo Equipment, Collectibles...
- Electronics
TVs, Cameras, Computer Add-Ons, DVD Players...
- Toys & Games
Toys for Grownups, Games & Puzzles, Action Figures, Baby Toys...
- Free E-Cards
Daily Cards, Hi, Love, Friendship...
- ...And More
Women's Apparel, Garden & Patio, Software, Sporting Goods...

Community
- Purchase Circles

Gift Services
- Gift Certificates
- Special-Occasion Reminders
- Shipping

Need Help?
- Help Desk
- Site Guide
- Shipping Policies
- Send Us E-mail
- Our Privacy Policy

More to Explore
- Join Associates
- E-mail Recommendations
- Pets.com
- drugstore.com
- Amazon.co.uk
- Amazon.de

What are they listening to in Austin, Texas? Explore Amazon.com Purchase Circles.

Amazon.com 100 Hot Books
Updated Hourly
1. Harry Potter and the Prisoner of Azkaban by J. K. Rowling
2. Harry Potter and the Sorcerer's Stone by J. K. Rowling
3. Harry Potter and the Chamber of Secrets by J. K. Rowling

More Hot Books

Today in Music
Free song downloads.
Music Top Sellers
1. Western Wall: The Tucson Sessions Linda Ronstadt, Emmylou Harris
2. Fly Dixie Chicks
3. Supernatural Santana

More Music Top Sellers

Today in Video
Pre-order *The Matrix* on DVD.
VHS Top Sellers
1. Living Yoga: A.M./P.M. Yoga for Beginners Set
2. There's Something About Mary (Special Edition)
3. Tae-Bo Workout: Instructional and Basic

More VHS Top Sellers
DVD Top Sellers

* "The New York Times" is the registered trademark of The New York Times Company, which is not affiliated with and does not sponsor or endorse the services of Amazon.com.

Text Only

1-Click Settings | Shopping Cart | Your Account | Help
About Amazon.com | Join Our Staff
Copyright and disclaimer © 1996-1999, Amazon.com, Inc.

FIGURE 1 Snapshot of Amazon.com (1998)

2.2 | Analysis metrics

We analyzed multiple quantitative metrics: (a) HTML metrics, (b) screenshot size, and (c) color metrics. We counted the number of images via the img-tag, the number of hyperlinks via the a-tag and the overall amount of words in the HTML. Another metric is the size of the screenshots measured in kilobytes after transforming

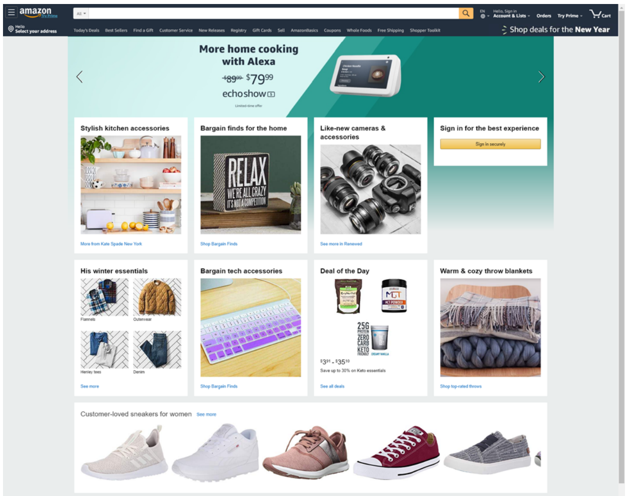

amazon
Select your address
Shop deals for the New Year
More home cooking with Alexa
$89.99 $79.99
echoshow
Stylish kitchen accessories
Bargain finds for the home
Like-new cameras & accessories
Sign in for the best experience
Sign in securely
His winter essentials
Bargain tech accessories
Deal of the Day
Warm & cozy throw blankets
Customer-loved sneakers for women

FIGURE 2 Snapshot of Amazon.com (2020)

TABLE 3 Overview of the used RGB color ranges

Color	Lower limit (R,G,B)	Upper limit (R,G,B)
Red	(140, 0, 0)	(255, 56, 50)
Green	(0, 170, 0)	(130, 255, 70)
Blue	(0, 0, 145)	(60, 115, 255)
Yellow	(230, 220, 0)	(255, 255, 55)
White	(240, 240, 240)	(255, 255, 255)
Black	(0, 0, 0)	(25, 25, 25)

them to PNG-files. This is an established metric to measure visual complexity and has been shown to correlate with human perception of complexity (Purchase et al., 2012); the larger the more complex an image. We calculated the amount of the base colors red, green, blue, yellow as well as white and black via openCV.[5] Table 3 gives an overview of the RGB-sections we included.

3 | RESULTS

We averaged the number of image-tags, hyperlink-tags and the text per year (see Figures 3–5).

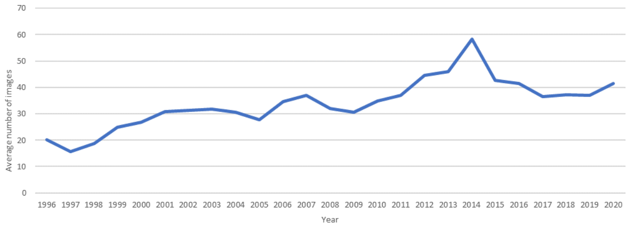


FIGURE 3 Average number of images per year

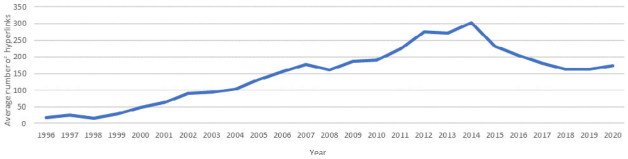


FIGURE 4 Average number of hyperlinks per year

For the number of images and hyperlinks we can identify a steady increase up until 2014 followed by a substantial decrease. A similar development can be found for text up until 2016.

Figure 6 shows the development of the average visual complexity as measured via the filesize.

For this metric, we identified a steady increase until 2020, with the highest leap in the time span from 2009 until 2014.

The color analysis shows that black and white are the most dominant colors since white is the general background color and black the basic font color. Figure 7 illustrates the average proportion of black and white among the websites and shows that most websites consist to around 80% percent of white. However, we observe a small decrease up until now.

While the overall proportion is much lower, we identified red and blue as the most used among the base colors, however without a consistent trend (Figure 8). Beginning with 2013, we can see a more diverse distribution among our analyzed colors. Nevertheless, without a striking trend or development. Please note that the proportions are overall very low, so the significant usage of very popular websites of one color for a year can lead to strong manifestations of this color – which is especially the case for years we do not have many snapshots for (1996–2003).

4 | DISCUSSION

We identified a steady increase of HTML-tags up until 2014 and a remarkable decrease afterwards. A similar but less strong trend is found for the number of words. We hypothesize that these results represent a trend in web design to include more content up until 2014, which is

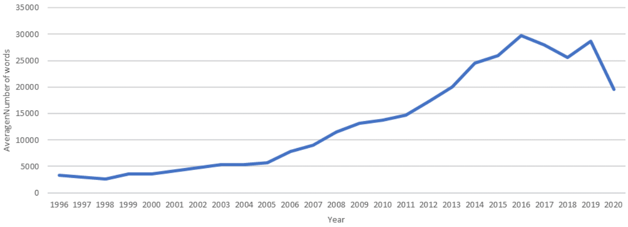


FIGURE 5 Average number of words per year

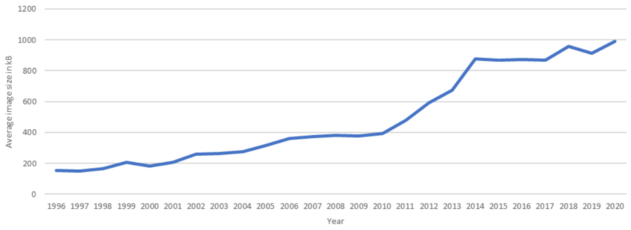


FIGURE 6 Average size of PNG-files per year

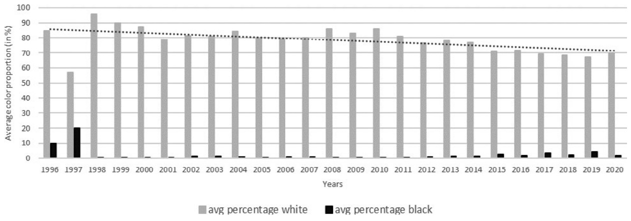


FIGURE 7 The average proportion of black and white among websites

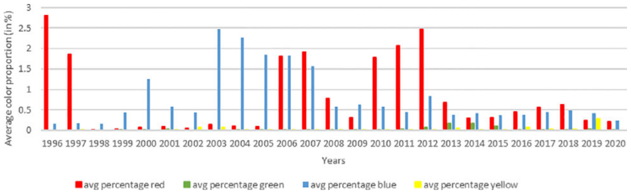
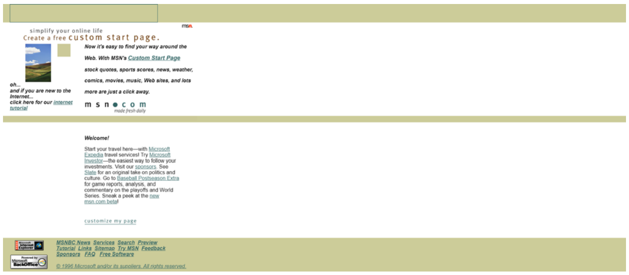


FIGURE 8 The average proportion of red, blue, green and yellow per year



msn.

simplify your online life
Create a free custom start page.

Now it's easy to find your way around the Web. With MSN's Custom Start Page stock quotes, sports scores, news, weather, comics, movies, music, Web sites, and lots more are just a click away.

oh...
and if you are new to the Internet...
click here for our internet tutorial

m s n ● c o m
made fresh daily

Welcome!

Start your travel here—with Microsoft Expedia travel services! Try Microsoft Investor—the easiest way to follow your investments. Visit our sponsors. See Slate for an original take on politics and culture. Go to Baseball Postseason Extra for game reports, analysis, and commentary on the playoffs and World Series. Sneak a peek at the new msn.com beta!

customize my page

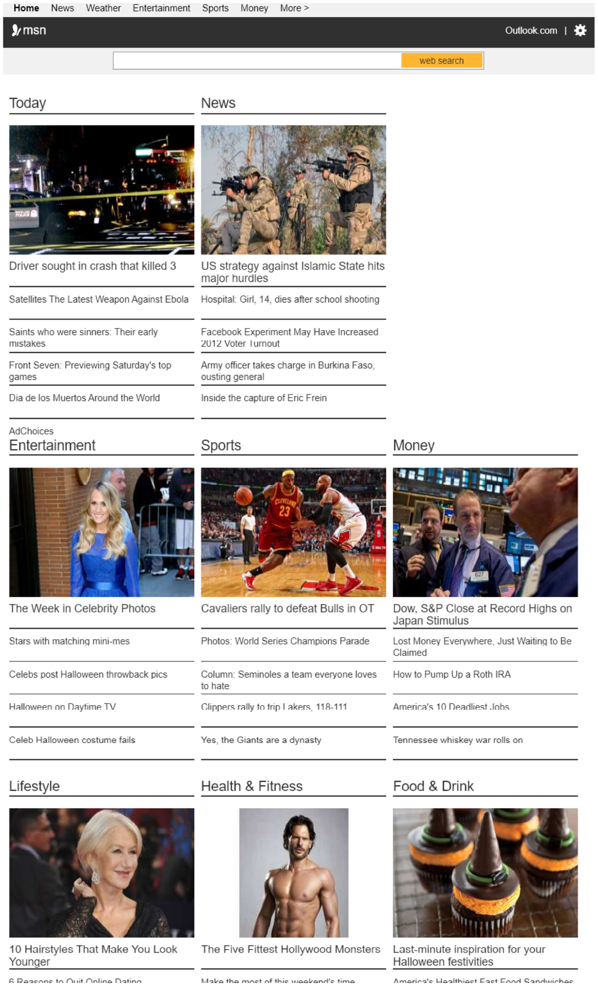
MSNBC News Services Search Preview
Tutorial Links Sitemap Try MSN Feedback
Sponsors FAQ Free Software

© 1996 Microsoft and/or its suppliers. All rights reserved.

FIGURE 9 Snapshot of Msn.com (1996)

certainly connected to increased technical possibilities. Snapshots of msn.com in our corpus are representative of this trend (Figures 9 and 10).

After 2014, this trend is followed by a tendency toward more minimalistic designs. This is opposite to the steady increase of visual complexity. The reason for the development of this metric, however, might also be the possibility for designers to include more complex images and graphs with the advent of bandwidth. While the usage of black and white is dominant on websites until now, we identified a higher usage of red and blue in early days and a more



Today

Driver sought in crash that killed 3

Satellites The Latest Weapon Against Ebola

Saints who were sinners: Their early mistakes

Front Seven: Previewing Saturday's top games

Dia de los Muertos Around the World

News

US strategy against Islamic State hits major hurdles

Hospital: Girl, 14, dies after school shooting

Facebook Experiment May Have Increased 2012 Voter Turnout

Army officer takes charge in Burkina Faso, ousting general

Inside the capture of Eric Frein

AdChoices

Entertainment

The Week in Celebrity Photos

Stars with matching mini-mes

Celebs post Halloween throwback pics

Halloween on Daytime TV

Celeb Halloween costume fails

Sports

Cavaliers rally to defeat Bulls in OT

Photos: World Series Champions Parade

Column: Seminoles a team everyone loves to hate

Clippers rally to trip Lakers, 118-111

Yes, the Giants are a dynasty

Money

Dow, S&P Close at Record Highs on Japan Stimulus

Lost Money Everywhere, Just Waiting to Be Claimed

How to Pump Up a Roth IRA

America's 10 Deadliest Jobs

Tennessee whiskey war rolls on

Lifestyle

10 Hairstyles That Make You Look Younger

6 Reasons to Quit Online Dating

Health & Fitness

The Five Fittest Hollywood Monsters

Make the most of this weekend's time

Food & Drink

Last-minute inspiration for your Halloween festivities

America's Healthiest Fast Food Sandwiches

FIGURE 10 Snapshot of Msn.com (2014)

diverse color usage beginning 2013. Our color analysis is however very limited since we neglect a wide range of other colors.

We want to continue our research by including other quantitative metrics like alternatives for visual complexity and other colors and by analyzing the websites by category. We plan to increase the corpus, especially with websites that were popular in the years 1996–2010 and are not nowadays to get a more representative sample of these times. We pursue a mixed methods approach and want to integrate qualitative analysis of a subset of the corpus to get a better understanding of design developments.

ENDNOTES

[1] https://en.unesco.org/themes/information-preservation/digital-heritage

[2] https://archive.org/

[3] https://archive.org/web/

[4] https://www.alexa.com/

[5] https://opencv.org/

REFERENCES

Brügger, N. (2012). When the present web is later the past: Web historiography, digital history, and internet studies. *Historical Social Research/Historische Sozialforschung*, *37*(4(142)), 102–117.

Jørgensen, A. H., & Myers, B. A. (2008). User interface history. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems* (pp. 2415–2418).

Purchase, H. C., Freeman, E., & Hamer, J. (2012). An exploration of visual complexity. In *International Conference on Theory and Application of Diagrams* (pp. 200–213). Berlin, Heidelberg: Springer.

How to cite this article: Schmidt T, Mosiienko A, Faber R, Herzog J, Wolff C. Utilizing HTML-analysis and computer vision on a corpus of website screenshots to investigate design developments on the web. *Proc Assoc Inf Sci Technol*. 2020;57:e392. https://doi.org/10.1002/pra2.392