Int J Digit Libr (2018) 19:95–106
https://doi.org/10.1007/s00799-016-0202-6

The colors of the national Web: visual data analysis of the historical Yugoslav Web domain

Anat Ben-David[1] · Adam Amram[2] · Ron Bekkerman[2]

Received: 1 March 2016 / Revised: 30 November 2016 / Accepted: 5 December 2016 / Published online: 18 December 2016

Abstract This study examines the use of visual data analytics as a method for historical investigation of national Webs, using Web archives. It empirically analyzes all graphically designed (non-photographic) images extracted from Websites hosted in the historical .yu domain and archived by the Internet Archive between 1997 and 2000, to assess the utility and value of visual data analytics as a measure of nationality of a Web domain. First, we report that only 23.5% of Websites hosted in the .yu domain over the studied years had their graphically designed images properly archived. Second, we detect significant differences between the color palettes of .yu sub-domains (commercial, organizational, academic, and governmental), as well as between Montenegrin and Serbian Websites. Third, we show that the similarity of the domains' colors to the colors of the Yugoslav national flag decreases over time. However, there are spikes in the use of Yugoslav national colors that correlate with major developments on the Kosovo frontier.

Keywords Web archives · Color analysis · Visual data analytics · Yugoslavia · National domain · Internet Archive

1 Introduction

Since the early 2000s, national libraries preserve countries' national digital heritage through archiving Websites registered under each Country Code Top Level Domain (ccTLD) [38,40]. Although the *Internet Archive* preserves both national and nationality-agnostic domains, national libraries invest substantial resources into increasing the scope, completeness, and quality of their Web archives. The appropriation of portions of the Web for national heritage preservation purposes thus emphasizes the importance of national self-determination as part of the (relatively short) history of the Web, as well as of the prospects of the future memory of the Web.

Although the early research used the Internet Archive to analyze social and political phenomena of national relevance, such as political campaigns and the September 11 attacks in the US [43], the emergence of national Web archives has triggered scholarly interest in using the archived Web as the primary source for studying the Web histories of nations. On the one hand, national Web archives are rich sources of information for analyzing historical events of the past two decades. From this perspective, Web archives attract the interest of historians, social scientists, and media researchers who attempt to answer questions that tie the history of the Web with the histories of nations. On the other hand, national Web archives constitute one of the first born-digital, paperless archives of national scale, and importance. From that perspective, archivists and library scientists are interested both in the archiving process itself, as well as in facilitating access to their scholarly research. Although methodologies for the study of *textual content* of (non-Web) archives can be fairly straightforwardly adapted to studying Web archives, Web archives are unique primary sources, whose specific characteristics extend beyond textual content. Web archives

✉ Anat Ben-David
anatbd@openu.ac.il; anatbd@gmail.com

Adam Amram
adam.amram@gmail.com

Ron Bekkerman
ronb@univ.haifa.ac.il

[1] The Open University of Israel, Ra'anana, Israel

[2] University of Haifa, Haifa, Israel

preserve a medium that is both ephemeral and multimodal, which necessitates the creation of a new analytical toolset to facilitate their scholarly use [16].

This study complements the exploration of national Web archives by investigating a Web domain that no longer exists and does not have a national Web archive. It builds upon the previous research on the fate of the ccTLD of former Yugoslavia (.yu), which operated on the Web from 1989 until its removal in 2010 [6]. Because both the country and its national domain are no longer a part of the live Web, the historical research of the .yu domain could be seen as a rather abnormal and exceptional case study, where the Internet Archive is the only entity that documented evidence of its existence in the past. Despite its abnormality, or rather due to its abnormality, the Yugoslav Web domain makes an interesting case for the study of nations using Web archives.

In our previous research [6], we reconstructed the historical .yu domain from the Internet Archive and demonstrated the utility of hyperlink network analysis as an effective method for analyzing the structure and temporal dynamics of the long vanished domain. Our current research utilizes the data from the reconstructed domain, while moving beyond the question of the feasibility of reconstructing the domain's timeline. We present a novel methodology for characterizing and exploring cultural elements in the history of a national Web. Specifically, we focus on the domain's archived images as units of analysis. We develop visual data analytics tools that enrich the scope of analyzing archived Web materials of national domains.

Arguably, empirical studies of national Web archives can be divided into structural analyses which examine their scope, completeness, and other technical elements (such as the hyperlinked structure of archived Websites or the distribution of file extensions and mime types of a national Web archive) [3,19], and historical studies that aim to draw significant social and cultural insights about the evolution of a particular Web domain [20,42]. Recently, structural and social studies of Web archives began to merge, resulting in the structural elements of the archive being treated as crucial analytical units from which historical, social, and cultural insights may be drawn [9,21,49].

This study brings together the structural, historical, and social approaches to Web archive research, by examining the extent to which structural analysis of images found in archived snapshots of the .yu domain can be used as units of analysis for both enriching the characterization of an archived national Web, as well as for drawing social and cultural implications of their evolution and history. In particular, we explore the ties between the domain's history and the evolution of the colors of the domain, during the early days of the Web (1997–2000), which—in the case of the .yu domain—coincided with political and economic instability, rising nationalism, and wars.

The goals of this study are twofold. First, we seek to develop visual data analytics as a means to identifying and characterizing digital nationality using Web archives. Second, we seek to theorize the ways with which the historical development of the Web plays a role in shaping cultural memory and digital cultural heritage. To meet these goals, our research is based upon a case study of the archived .yu domain and poses the following research questions:

RQ1: Does the .yu domain have typical color schemes and do they evolve over time?
We address this question by calculating the average color histogram of images found in each archived page of the .yu domain and analyze the changes in the domain's dominant colors over time.

RQ2: Is there a correlation between the domain Websites' primary colors and other characterizing elements of the domain (e.g., distribution of sub-domains and hyperlink structure)?
We draw temporal and topological diagrams of the Yugoslav Web and overlap them with the average colors of sub-domains or major Websites, resulting in colorful collages that provide a 30,000-feet view of the .yu domain, which allow a comprehensive study of the domain as a whole.

RQ3: Does the histogram of the national flag of Yugoslavia correlate with the colors of the domain and does the correlation change over time during the years of the Kosovo War?
We monitor the online usage of the national colors of Yugoslavia—the three colors of the flag of the Federal Republic of Yugoslavia (blue, white, and red—see Fig. 1, left), the union state between Serbia and Montenegro—over the studied period (1997–2000), to detect possible correlations between temporal changes in the colors of the domain and national events. To do so, we compute the proportion of the national colors within the major colors of all Yugoslav Websites. We draw a temporal diagram that shows the evolution of the relative 'distance' of the domain's colors from the national colors over time, per sub-domain.

RQ4: Given the topology of the Yugoslav domain as represented by the set of hyperlinks between its Websites, as well as the archived color palettes of those Websites, and considering the incompleteness of Web archives, can we infer the color schemes of those Websites whose images were not archived?

The phenomenon of topical locality of the Web [15] let us assume that Websites connected to each other with hyperlinks tend to discuss similar topics. However, can we assume that this implies a certain level of similarity in the visual palette of those Websites? If so, can we infer the visual palette of Websites that were not archived properly, while they still existed, now that their visual features are gone forever? To answer this question, we draw a network graph of the top 106 most interlinked Websites in the .yu domain, and examine the

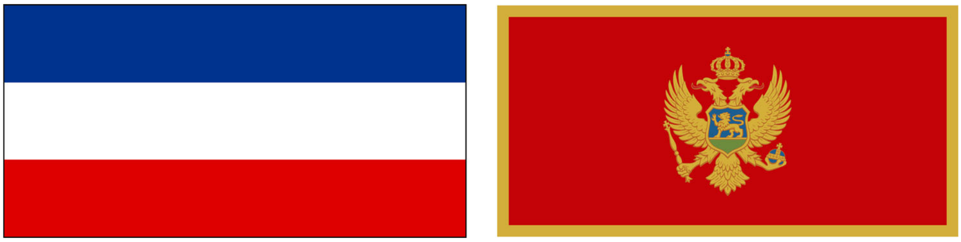

Fig. 1 Flag of the Federal Republic of Yugoslavia (*left*); flag of Montenegro (*right*)

extent to which the color palette of missing images can be inferred from the colors of neighboring nodes.

The following parts of this paper are structured as follows: we first provide a short overview of related work, then describe our proposed methodology, and present our findings: the analysis of the overall color characteristics of the .yu domain, the analysis of correlation between domain colors and its structural elements (hyperlink structure, sub-domain distribution), and the analysis of the dynamics of the national colors on the domain's Websites. A discussion on the applicability and limitations of our methodology concludes this paper.

2 Related work

In recent years, analysis of archived Web data has expanded its scale from a single Website or a selection of interlinked Websites to the study of the complete national Web archive as a unit of analysis [3,37]. For example, Hale et al. [21] used a snapshot of the .uk domain archived by the Internet Archive between 1996 and 2010 to map the structural evolution of the .uk domain and to characterize historical linking practices of British universities. Similarly, the entire national Web archive of Denmark has been analyzed to study the history of the Danish Web, using metrics, such as the size of Websites, their geo-location, and hyperlinked structure [9]. Researchers also analyzed the Dutch Web archive to assess the extent to which evidence of unarchived pages can be retrieved [27].

The Web history of countries without an official national Web archive can be retrieved from the Internet Archive, although studies show substantial differences in the archival coverage of different national Webs [2,46]. A recent study of the Yugoslav domain, which operated from 1989 until it was deleted from the Internet's root zone in 2010, reconstructed its historical past from the Internet Archive using a hyperlink discovery method [6].

While Web archives are gaining traction over recent years as objects of study, the literature expresses growing concerns about their incompleteness. Although the premise is that it is impossible to archive the entire Web, Web archives are incomplete not only due to the limited update frequency or selection policy, but also due to the loss of important Website elements (such as linked content, videos, images, or Flash) during the archiving process [8,11,24,29]. Such incompleteness is accompanied by missing contextual elements on the archiving process itself, e.g., insufficient metadata about the selection method, archiving frequency, and time of archiving, or 'entity level analytics' that researchers need to comprehensively analyze the content of a Web archive [37,48].

Recent research has paid attention to the incompleteness of Web archives by estimating the volume of Web archives' coverage [1,3], and by evaluating the extent to which it is possible to deduct evidence about moved or unarchived pages [27,31]. For these reasons, researchers active in the field of digital humanities argue that at their current state, Web archives are not widely used, because they do not lend themselves to performing 'distant reading' or applying other large-scale analytical methods for studying born-digital corpora [32,36,48].

Although recent effort is focused on implementing textual search to improve access interfaces to Web archives [7,14], the application of big data analytics and distant reading techniques on large corpora of archived Websites should not be limited to text, but rather consider other archived multimodal content and formats, such as images, graphs, and videos [10,26,48].

Among the multimodal elements of Web archives, color is rather abstract. Although all people experience colors, there is not always agreement about what people actually perceive [18]. From a technical perspective, colors can be represented and manipulated by different color spaces, or coordinate systems. One of the common color spaces is RGB. The RGB color space consists of three components: red, green, and blue. Any combination of these three components will produce a resultant color [50]. Another color space is the CIE-Lab. The CIE-Lab color space includes all perceivable colors and is based on translation into distinctions between light and dark (L-axis), red and green (a-axis), and blue and yellow (b-axis) [22].

Since colors are organized in a coordinate system, or color space, differences between colors can be measured as distance. It is estimated that as many as 20 different formulae were used in various industries in various parts of the world to calculate color difference before the International Commission on Illumination (CIE), a technical, scientific, and cultural organization that provides basic standards and procedures in the fields of light and lighting, recommended using the CIE-Lab color space and the delta-E formula to measure differences between colors [12,41].

Studies on visual data analytics focus on four levels of information while dealing with image mining: pixel level; object level; semantic concept level; and pattern and knowledge level [51]. Pixel level is the lowest level in image mining at which primitive image features, such as color, texture, and edge information, are extracted. Color is the most commonly used feature because of its strong correlation with the underlying image objects [45,47].

Historically, the study of images and color of Websites dates back to the early days of the Web and primarily relates to the early research on Web design, putting an emphasis on the role of Websites as cultural markers that may influence users' behavior or distinguish between different cultures [4,5,13,34,39]. An early study of the Yugoslav Web, for example, used media-richness theory to tie hyperlinking practices, visual representation, and geographic placements of the former Yugoslavia on the Web [28]. While the early research practices applied manual classification methods for evaluating Websites' images and colors, recent research focuses on automated image mining as a common practice in computational social sciences and digital humanities, where the aesthetic and technical qualities of large corpora of digital images are analyzed as a means to offer social, cultural, and political insights [23]. In this sense, and despite the difficulties mentioned above in applying distant reading techniques on Web archives, Web archives add unique temporal attributes to visual analytics. Milligan [36], for example, discusses the disappearance of color from Canadian federal Websites: "moving through the preserved Websites is an interesting snapshot at how Web design standards changed. Color disappears, replaced by a more subdued coordinated federal Website standard (still largely in effect today)" (p. 37).

3 Method

3.1 Data collection

The demarcation of a national Web is a methodological challenge that involves Websites hosted both within and outside a country's ccTLD [19]. Although the previous research confirmed that a considerable number of historical Yugoslav Websites were hosted under generic domains [6,35], the data presented in this analysis are limited to the .yu domain as we propose a method for the visual characterization of a ccTLD, which can be later applied to other ccTLDs. Therefore, our first step was to collect all the available data of the .yu domain. Using Python scripts, we crawled the .yu domain of the Internet Archive. Our crawler was initiated with a seed list of 16,758 known .yu URLs, which was constructed in the previous work [6]. For each URL in the list, the crawler retrieved all available snapshots of the URL content between the years 1997 and 2000, from the Internet Archive's Wayback Machine CDX Server.[1] Then, for each available snapshot, the crawler collected all available data. Each retrieved item (an HTML page, an image, a PDF file, etc.) was saved as a document in an Elastic Search[2] server. Collected items were saved with their textual and visual content, archival date, mime type, and marked as 'found'; items that were hyperlinked but were not archived by the Internet Archive were saved without their content and marked with the status 'not found'. For example, if one HTML page linked to an image, but the image could not be found in the Internet Archive, a document containing the link to the image and the status 'not found' was saved.

We accessed the content of 2333 Yugoslav Websites archived between the years 1997–2000. From the archived content, our crawler retrieved 936,549 items (Webpages, images, documents, etc.), out of which a total of 97,845 images were associated with Yugoslav Websites.

Note that for some periods of time, there are no snapshots of the Yugoslav Web in the Internet Archive. Namely, two periods (August–October 1998 and June–September 1999) are not covered at all. Although we do not have strong evidence to account for possible reasons for the outage, we suppose that the Web archiving gaps were consequences of the NATO campaign during the Kosovo War (probably caused by electricity outages or the Internet disruptions in the Federal Republic of Yugoslavia during those time periods) [30,33].

3.2 Extraction of dominant colors

To extract the dominant colors, a distinction was made between graphically designed images (e.g., logos, background images, buttons, and other parts of the graphical user interface) and photographs. Our rationale was that the color palette of a photograph was not always under control of a Website designer, especially in the early days of the Web. Hence, to analyze the usage of colors in the Yugoslav Web, we had to exclude photographs from our consideration.

To separate photographs from other images, we used the following heuristic. Each image was represented by its color histogram. We assumed that a histogram of a photograph is significantly richer than the one of a designed image, which would contain a small number of dominant colors. We inferred that an image was graphically designed if it had at least two colors, each of which covering at least 8% of

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

[2] Elastic Search is a distributed, open source search, and analytics engine, designed for horizontal scalability, reliability, and easy management. It combines the speed of search with the power of analytics via a sophisticated, developer-friendly query language covering structured, unstructured, and time-series data.

Table 1 Number of graphically designed images, per sub-domain

Sub-domain	1997	1998	1999	2000	Total
ac.yu	233 (72)	302 (158)	544 (160)	544 (148)	1623 (538)
cg.yu	0 (0)	82 (38)	209 (106)	739 (154)	1030 (298)
co.yu	198 (168)	1326 (735)	5399 (1946)	15,471 (3857)	22,394 (6706)
edu.yu	0 (0)	24 (24)	82 (40)	139 (36)	245 (100)
org.yu	57 (32)	232 (130)	917 (373)	3237 (635)	4443 (1170)
.yu	71 (29)	109 (36)	212 (67)	599 (88)	991 (220)
Total	559 (301)	2075 (1121)	7363 (2692)	20,729 (4918)	30,745 (9032)

The numbers in parentheses indicate the number of unique graphically designed images

the image area. Images that did not meet this criteria were inferred as photographs. This heuristic worked extremely well in practice; manual examination of a portion of the results did not reveal one single mistake. About two thirds of the overall number of images were recognized as photographs and filtered out.

After filtering the photographs out, we extracted the dominant colors for each sub-domain per month. We considered six major sub-domains of the .yu domain:

1. .ac.yu: the sub-domain for academic institutions (such as universities).
2. .cg.yu: Montenegrin sub-domain.
3. .co.yu: commercial, for-profit sub-domain.
4. .edu.yu: the sub-domain for educational institutions (such as primary schools and high schools).
5. .org.yu: organizational, non-profit sub-domain.
6. .yu: sub-domain of governmental institutions and Internet Service Providers.

Table 1 shows the statistics on the number of graphically designed images found, per year, and sub-domain. The numbers in parentheses indicate the number of unique graphically designed images.

We can see a sharp increase in the use of graphically designed images over the years. In particular, the commercial sub-domain (.co.yu) experienced an increase of 7713% in the usage of graphically designed images between the years 1997–2000. The non-profit sub-domain (.org.yu) followed .co.yu. This can be easily explained by the rise of the commercial value of the Web: the years of 1997–2000 were the early years of e-commerce and e-marketing, when the importance of sleek Web design was recognized by Website owners. The transformation from the early, text-only designs to much more impressive graphical forms resulted in a substantial increase in the amount of graphically designed primitives used on the Web.

We used Python scripts to create an image collage from all unique non-photographic images. We created collages for the following levels: (a) each month over the whole .yu domain; (b) each month and each .yu sub-domain (e.g., .org.yu, .ac.yu, etc.); and (c) each domain over all years (e.g., http://www.spo.org.yu, http://www.skypass.co.yu, etc.). Figure 2 (left) shows part of the collage that was composed from images found in http://www.motomat.co.yu. To improve representation and performance, we resized the images used for building the collage to a maximum of 15 × 10 pixels. Since the Internet Archive may hold the same image objects across many snapshots of a page, which can overweight the color palette, we performed a deduplication procedure. In the process of making the collage, we used a hash function to generate a unique signature of the images. The hash function guarantees that if two images are identical, it will generate the same signature. Therefore, if a signature of an image was already used in a collage, the image was ignored. For each collage, we applied the K-means clustering algorithm to create five clusters of colors of the collage's pixels. The dominant colors of the collage were then assumed to be the centroids of the three largest clusters (see Fig.2, right).

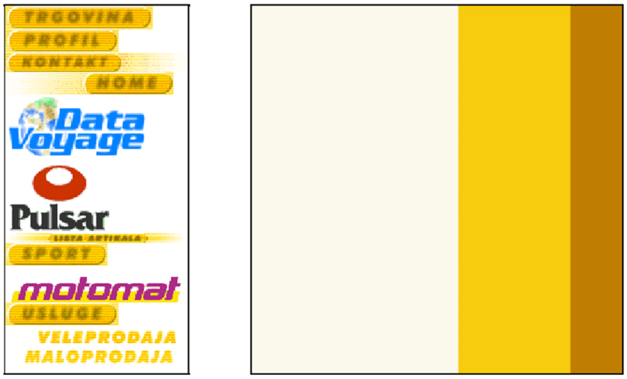

Fig. 2 Collage from extracted images of http://www.motomat.co.yu (*left*); three dominant colors of http://www.motomat.co.yu (*right*)

3.3 Similarity to the national flag

We compared the three dominant colors of each sub-domain and the three colors of the flag of the Federal Republic of Yugoslavia (blue, white, and red). Specifically, the RGB values of the flag colors were taken as an average of a few images of the Yugoslav flag, which were: (a) Blue (R: 1, G: 56, B: 147), (b) White (R: 255, G: 255, B: 254), and (c) Red (R: 220, G: 0, B: 0) (see Fig. 1, left). Each of the flag colors was compared to each dominant color extracted as explained above. As a distance measure between two colors, we used the International Commission on Illumination (CIE) delta-E function from 1994 [44]. The delta-E function yielded a single number representing the distance between two colors. If the delta-E value was below 20, the sub-domain color was considered close to the flag color; otherwise, it was considered far and thus irrelevant. For each month and sub-domain, an average similarity score was calculated between the dominant colors of the sub-domain and the colors of the Yugoslav flag.

3.4 Network of colors and missing images

From the collected data, we created a graph of Yugoslav Websites that hyperlinked with each other. As nodes in the network, we show images representing the three most dominant colors of each Website, which were obtained using the same technique described before. If visuals of a specific Website were not archived properly in the Internet Archive, the node was colored all-grey.

4 Findings

4.1 RQ1: distribution of colors in the Yugoslav Web over time

To answer the question whether the .yu domain has typical color schemes and do they evolve over time (RQ1), we calculated the average color histogram of the images found in each archived page of the .yu domain, and analyzed the changes in the domain's dominant colors over time (see Fig. 3). As can be seen in Fig. 3, although the number of designed images increases over time, the diversity of the domain's typical color schemes decreases. While between 1997 and 1999, the typical colors of the domain are heterogeneous, in 2000, there is a noticeable convergence of colors around the shades of light blue, dark blue (black), and light grey—colors that are generally associated with the e-commerce look [39].

4.2 RQ2: distribution of colors per sub-domain

When the distribution of the domain's color palette is broken into sub-domains (see Fig. 4), a greater affinity can be noticed between colors and sub-domains, than between colors and time. That is, the color palettes of each sub-domain are visibly different from each other. For example, brown is a frequent color in the organizational domain .org.yu, but not in commercial Websites hosted under the .co.yu sub-domain.

In contrast, the sub-domain colors do not change dramatically over time. Here, we see a clear separation between two time periods: the period of Web immaturity (prior to November 1998) and the period of Web maturity (starting from November 1998). During the first period, a high diversity of colors is evident. Around November 1998, two new sub-domains are added (the Montenegrin sub-domain and the high education sub-domain). From 1998 onwards, the main color palettes of the sub-domains become less diverse, while the top three colors do not significantly change over time. This is especially apparent in the commercial sub-domain, in which the palette of blue, black, and grey does not change from 1998 to 2000.

A possible explanation for the shift from a diverse color palette to a more homogeneous one would be the evolution of new Web authoring tools (such as the introduction of Cascading Style Sheets in 1998) and the standardization of style in Web design [17]. In addition, as the Web had passed the commercialization transformation by the end of 1998, many commercial Websites had defined their value proposition and no longer felt a need for experimentation. On the other hand, there might be a completely different explanation for the palette stabilization, which has to do with the exceptional history of the .yu domain: between 1994 and 2000—years marked by wars, sanctions, and regional instability—the .yu domain was voluntarily maintained by the University of Belgrade. However, registration of new addresses had been limited to one Website per institution, due to the lack of financial resources and of governmental support [33]. Our previous examination of the evolution of the hyperlinked structure of the .yu domain confirms the stagnation of the development of the domain until the end of the Milosevic regime in 2000 [6]. Such stagnation is thus also apparent in terms of the visual development of the domain over these years.

4.3 RQ3: similarity to the national colors

Differences in the color distribution of designed images of Websites hosted in different sub-domains are also found when examining the similarity of the images of the .yu domain to the colors of the flag of the Federal Republic of Yugoslavia (see Fig. 5, left). Our examination of the relative distance of the domain's monthly color palettes per sub-domain shows substantial differences between Yugoslav sub-domains, primarily between Serbian sub-domains and the Montenegrin sub-domain (cg.yu). The color palettes of the Montenegrin sub-domain is much less similar to the colors of the Yugoslav national flag, compared to the similarity of the majority of Serbian sub-domains. The organizational sub-domain demonstrates the highest accordance with the national colors.

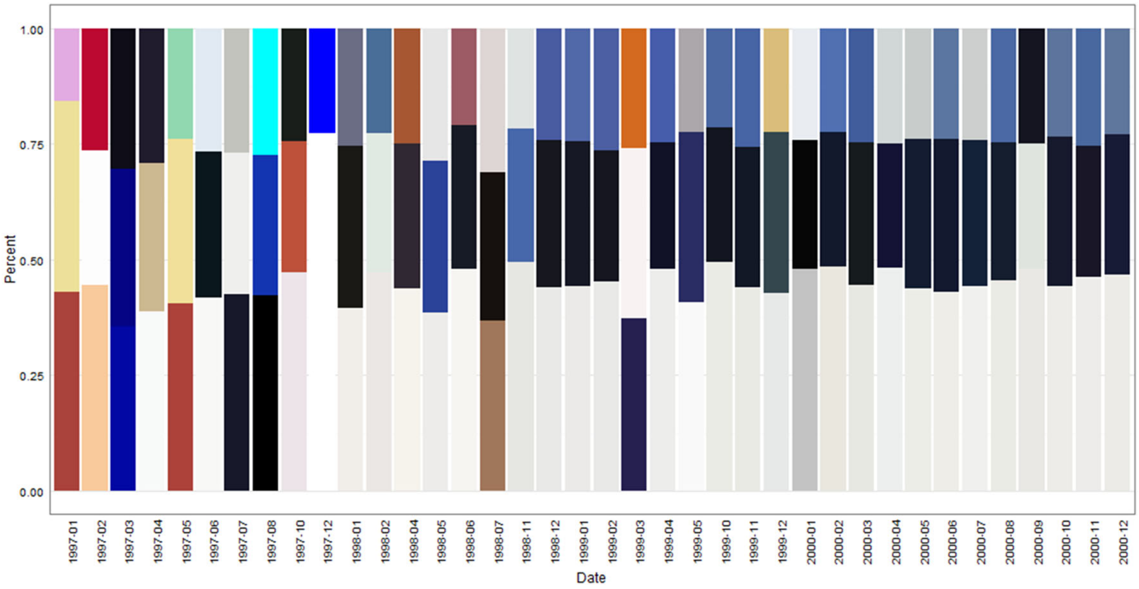

Fig. 3 Distribution of colors in the Yugoslav Web over time

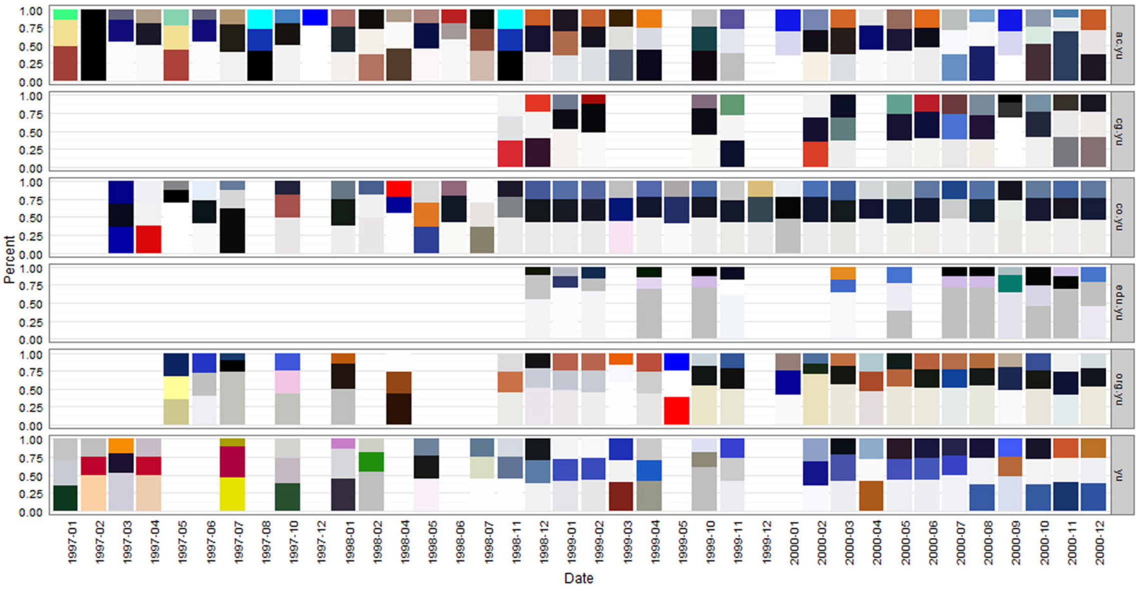

Fig. 4 Distribution of dominant colors in the Yugoslav domain, per sub-domain

We can see in Fig. 4 that there were several months in which the national colors were dominant in different sub-domains (April 1998 in the commercial sub-domain; May 1999 in the organizational sub-domain; March 1999 and November 2000 in the governmental sub-domain). Notably, the Montenegrin sub-domain (.cg.yu) had none, while the dark red color is more typical for that sub-domain, which may refer to the Montenegrin national flag (see Figs. 1, right, 4).

An examination of the temporal changes in the similarity of the domain's colors to the colors of the national flag shows that there is an overall decrease in similarity to the colors of the flag over time (see Fig. 5, right), which might correspond to the increasing apathy of the public towards their national identity as the war progressed and the union state got closer and closer to breaking down into separate states. However, we see some spikes in using the national colors, for example, in April 1999 during the time of the NATO bombing of Yugoslavia.

4.4 RQ4: structural holes in the Yugoslav Web palette

Our final analysis relates to the problem of incompleteness of Web archives, and pays particular attention to the extent to which images of the early Web have properly been archived

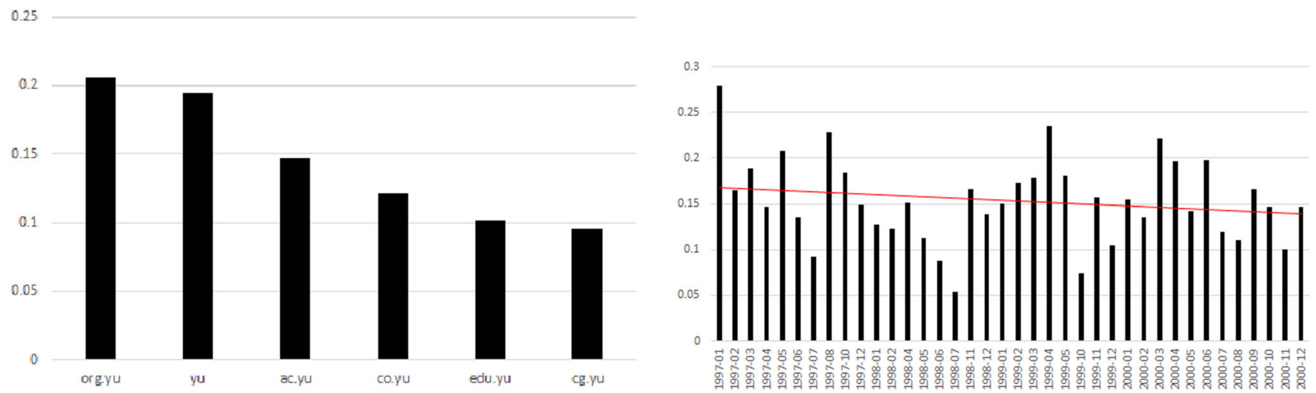

Fig. 5 Average score of similarity to the three colors of the flag of the Federal Republic of Yugoslavia: per sub-domain (*left*) and over time (*right*)

Table 2 Number of .yu images existing or missing in the Internet Archive by year and image type

Year	Existing		Missing	
	GIF	JPEG	GIF	JPEG
1997	1103	154	2338	0
1998	5143	2632	16, 334	0
1999	15,680	8023	55,462	8
2000	42,550	22,705	22,705	134

by the Internet Archive. An initial examination of our data reveals that compared to the JPEG image file type, GIF image files are particularly prone to incomplete archiving. During the studied period, as many as 60% of hyperlinks to GIF images lead to missing files, compared to only 0.4% of the JPEG images (see Table 2). Since our analysis focuses on graphically designed images and not on photographs (which usually appear in the JPEG format [25]), this finding has a crucial effect on our ability to infer from the color palette of existing graphically designed images, regarding the missing images.

An exciting experiment would be to visualize the graph of hyperlinks within the Yugoslav domain, where the nodes in the graph represent Yugoslav Websites and are marked with the dominant colors of each Website. The edges in the graph are the hyperlinks between the Websites. Such a visualization is yet another analytical tool that would give a holistic view of the entire domain and its color scheme. Using Gephi[3]—the popular tool for graph visualization—we built such a graph for the entire Yugoslav domain. Unfortunately, the scale of the graph is prohibitive for its representation in print, so, for the sake of clarity, we had to downgrade the scope of the graph presentation to only the 106 most connected nodes.

Since many Yugoslav Websites were archived without their visual features (that is, as bare HTML files)—we color those Websites in plain grey. Then, the important questions we are aiming to answer are as follows. Are there structural holes in the archive of the Yugoslav domain? In other words, can we use hyperlink analysis to see Website aggregations or clusters, that were archived comprehensively, while others were archived sparsely, in terms of their visual features? Can we characterize those structural holes?

Figure 6 shows the results. It is quite striking that the Websites are clustered in an almost perfect correspondence to their sub-domains: the academic sub-domain in the right upper corner, the organizational sub-domain in the right lower corner, the Montenegrin sub-domain in the bottom to the left, and Internet Service Providers in the left upper corner. The bulk of the domain is located in the center, corresponding to the largest, commercial sub-domain (.co.yu).

It is interesting to see that some Websites are very well connected to other Websites in the Yugoslav domain, while the others are weakly connected. We analyzed some of the well-connected Websites (see Table 3) and found out that most of them were link aggregators[4] (or news aggregators). Indeed, link aggregators were popular in the 1990s, as they attracted many visitors (who fueled the advertising business)—before getting banned by major search engines.

Considering the archival coverage of the Websites' visual features, we can easily see from Fig. 6 that the cluster corresponding to the academic sub-domain was not covered at all. In contrast, the organizational sub-domain cluster and the .yu sub-domain were relatively well-covered. The central cluster (the commercial sub-domain) was partly covered: the larger Websites in the cluster (represented by a larger color histogram) were covered well enough; however, smaller Websites were poorly covered. We, therefore, conclude that structural holes in archiving visual features exist,

[3] https://gephi.org/.

[4] A link aggregator is a Website that does not provide much authentic content, but rather presents a categorized list of links to other Websites.

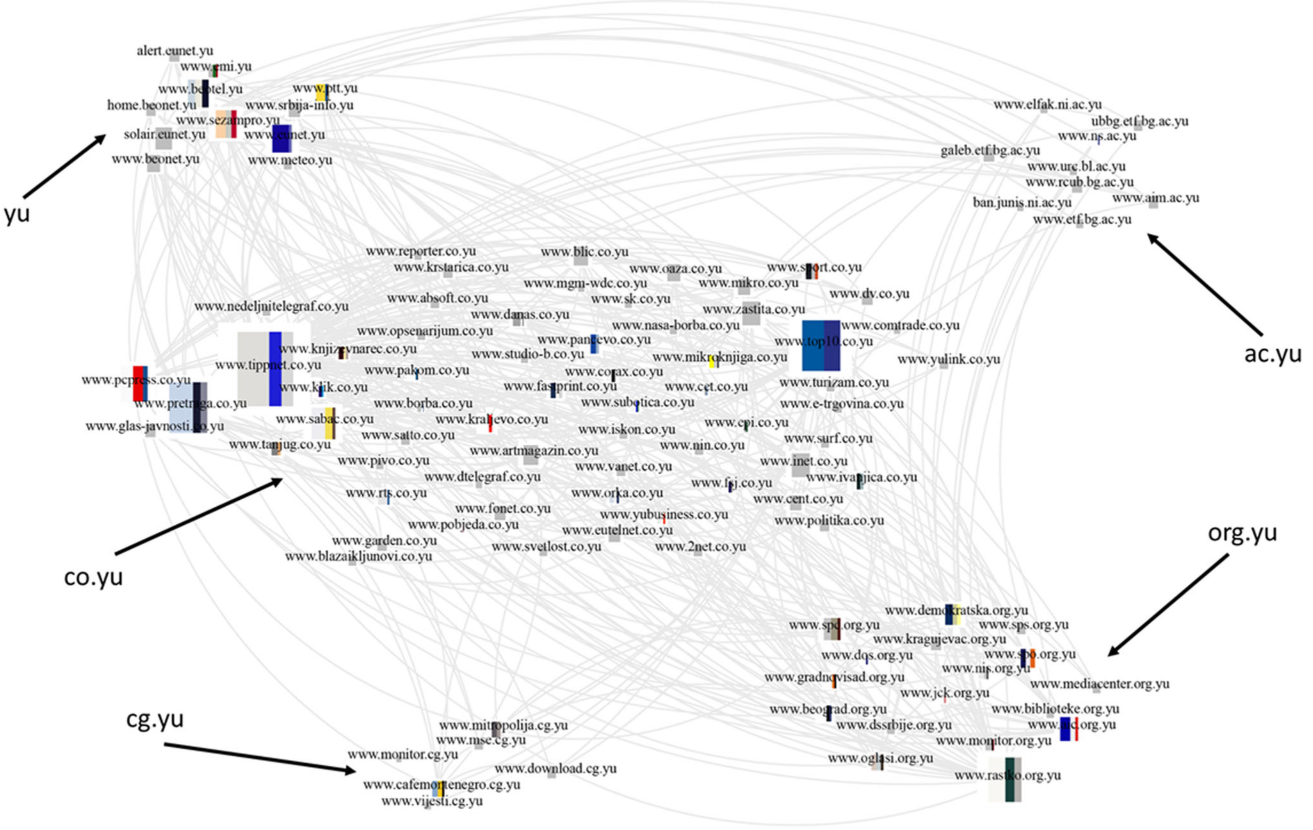

Fig. 6 106 most connected Websites organized as clusters according to their interconnectedness

Table 3 Most well-connected Websites in the .yu domain

http://www.tippnet.co.yu	News aggregator
http://www.map.co.yu	Internet archive of Yugoslavia
http://www.rastko.org.yu	Online library of Serbian culture
http://www.eunet.yu	Internet Service Provider
http://www.top10.co.yu	Ratings of everything
http://www.pretraga.co.yu	Link and news aggregator
http://www.network.cg.yu	Montenegrin link aggregator
http://www.pcpress.co.yu	Computer magazine
http://www.sezampro.yu	News and Web-related materials
http://www.betonjerka-al.co.yu	Internet Service Provider
http://www.infotrend.co.yu	Web design agency

indeed, where larger Websites are likely to be archived better than smaller Websites.

Figure 7 shows a close-up view of the organizational cluster from Fig. 6. It is evident that closely interlinked Websites have fairly different visual palettes. Coming back to RQ4, which dealt with the topical locality of the Web, it is quite clear that the topical locality phenomenon holds in the Yugoslav Web, as Websites that are similar in topic (e.g., non-profit organizations) are connected to each other. On the other hand, the fact that the majority of the graphically designed images in the data set were not properly archived, and that closely interlinked Websites with archived images do not seem to have similar color histograms, makes it difficult to determine whether or not Websites related in topic share similar visual palettes.

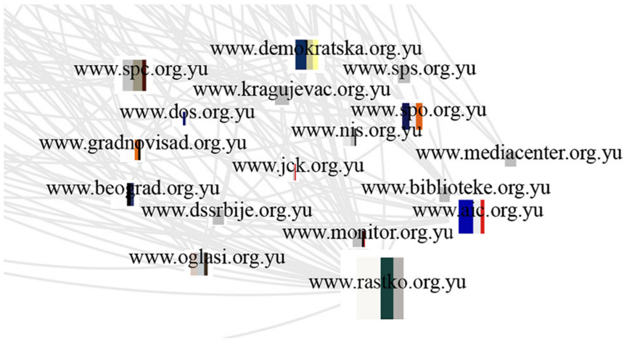

Fig. 7 Zoom onto the organizational Web of Yugoslavia

However, there is one color palette that appears to repeat itself between different Websites—this palette consists of the blue, white, and red colors, which are the colors of the Yugoslav flag! Although among the top 106 Websites presented in Fig. 6, only 46 (43.39%) of the Websites were

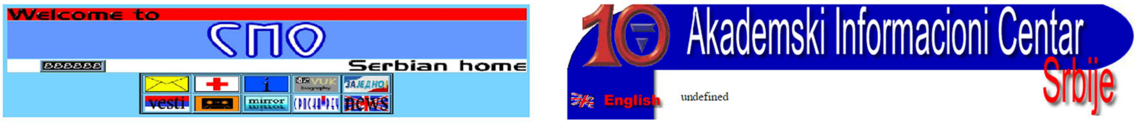

Fig. 8 Actual view of http://www.spo.org.yu (*left*, archived on June 1, 1997) and http://www.aic.org.yu (*right*, archived on April 28, 1999). The Yugoslav national color palette is apparent

archived with images, almost half of the Websites that were archived with graphically designed images (21, 45.6%) had at least one Yugoslav flag in the logos.

Figure 8 shows two such Websites, as archived with a gap of almost 2 years from each other. It is remarkable that both Websites have a very strong national appearance as their palette matches the colors of the Yugoslav national flag.

5 Discussion

In analyzing the changes in the overall color distribution of the .yu domain in the studied period, we found that the Yugoslav Web became less colorful over time, and that the prevalence of the use of the national flag's colors in graphically designed images decreased during the Kosovo war. Such 'visual distant reading' of the domain as a whole may be considered as a new method for analyzing national Webs, which extends beyond the dependence on textual search as a gateway to retrieving and analyzing images. While we couple the decrease in the resemblance of the domain's colors to the colors of the national flag over time with the gradual disintegration of the Federal Republic of Yugoslavia, we also notice significant differences between the color palettes of the Serbian and Montenegrin sub-domains within the .yu domain, thereby confirming that color does play a significant role as a cultural marker of national Webs. Put differently, the Serbian and Montenegrin Websites already had a different visual appeal even when still operating under a Union State.

At the same time, the analysis of the domain's colors could not stand alone as a method for characterizing a national Web archive, and needs to be aided by other structural elements of the domain for further contextualization and demarcation. In the case of the .yu domain, we consider the distribution of sub-domains as a structural element that contributes to revealing significant differences between different types of Yugoslav Websites. However, when coupling sub-domain distribution with hyperlink analysis as a contextualizing structural element, we found that the clustering of the .yu domain around different sub-domains did not significantly correlate with Websites' color distribution. Although future research may determine whether hyperlink topology and color distribution are, indeed, unrelated, our inability to significantly determine whether interlinked Websites share similar visual patterns is largely effected by the fact that the majority of the images of the Websites archived in the early days of the Web were missing.

The automated, 'distant reading' method applied in this study on the use of color in archived Websites of the former Yugoslav domain assumes a quantitative, data-driven approach to the analysis of Web archives. While such approach enables large-scale analyses of the national Web as a whole, it does not consider symbolic uses or meanings of the use of color. This limitation reflects in our seemingly contradictory findings: on the one hand, our quantitative analysis shows a gradual decrease in the overall use of the national flag's colors over time. On the other hand, our qualitative analysis of 46 Websites shows that almost half of them contain at least one logo depicting the national flag, mentioning the country's name, or displaying background colors in the blue–white–red palette. With this regard, our quantitative approach might fail to detect the symbolic meaning of one logo of the national flag to the construction of the .yu domain as a national space, even though the overall color histogram of the Website uses a different palette. Together with that, our

Table 4 Summary of the main findings

Research question	Findings
RQ1: Does the .yu domain have typical color schemes and do they evolve over time?	The diversity of the domain's typical color schemes decreases over time (Sect. 4.1)
RQ2: Is there a correlation between the domain Websites' primary colors and other characterizing elements of the domain?	The color palettes of each sub-domain are visibly different from each other but do not change dramatically over time (Sect. 4.2)
RQ3: Does the histogram of the national flag of Yugoslavia correlate with the colors of the domain and does the correlation change over time during the years of the Kosovo War?	Despite an overall decrease in the similarity of the domain's palette to the national flag, we identify spikes of similarity around events related to the Kosovo War. The color palette of the Montenegrin sub-domain is different from Serbian sub-domains, and is less similar to the Yugoslav flag (Sect. 4.3)
RQ4: Does hyperlink topology aid in inferring the palette of missing images?	Closely interlinked Websites with archived images do not seem to have similar color histograms (Sect. 4.4)

qualitative study covered less than 10% of Websites whose graphics were archived—and could not possibly scale to the entire data set. Table 4 shows a summary of the research questions and findings.

6 Future work

The .yu domain makes an exceptional case for the study of national Web archives, as it is both a vanished domain that does not continue to operate on the live Web, as well as a domain that operated in the early days of the Web during a period of war, political instability and economic sanctions. In that sense, it would be difficult to compare the findings from this study with other national Webs during the same years. At the same time, future research may expand the visual data analytics methods developed in this research to conduct comprehensive comparative analysis of possible differences between several national Web archives. Such comparison may help understanding whether some of the phenomena identified in this research—such as the convergence of the national domain around similar colors—are cultural practices that are shared by other national Webs. In addition, future work may further elaborate the method presented in this study to also include photographic images, and to automate the analysis of their content.

7 Conclusions

Web archives are valuable and unique historical resources rich with multimodal elements that can be used to analyze both the history of the Web as well as social history using the Web as its primary source. Of the various elements that can be used as units of analysis—such as text, HTML, or hyperlinks—color is a rather abstract entity, whose attributes are difficult to quantify and assess. Nevertheless, our analysis of the colors of the .yu domain, as it was archived by the Internet Archive, shows that color plays a non-arbitrary role as a characterizing element of a national Web domain. In particular, the combination of color and other structural elements of the domain (distribution of sub-domains and hyperlink topology) reveals patterns that distinguish between Websites and yields a dynamic view of the changes in the visual appeal of the national domain over time.

References

1. Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How much of the web is archived? In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, pp 133–136. ACM (2011)
2. Alkwai, L.M., Nelson, M.L., Weigle, M.C.: How well are Arabic websites archived? In: Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries, pp. 223–232. ACM (2015)
3. AlSum, A., Weigle, M.C., Nelson, M.L., Van de Sompel, H.: Profiling web archive coverage for top-level domain and content language. Int. J. Digit. Libr. **14**(3–4), 149–166 (2014)
4. Badre, A.: The effects of cross cultural interface design orientation on World Wide Web user performance. Technical Report GIT-GVU-01-03, Georgia Institute of Technology. https://smartech.gatech.edu/handle/1853/3315 (2001)
5. Barber, W., Badre, A.: Culturability: the merging of culture and usability. In: Proceedings of the 4th Conference on Human Factors and the Web, vol. 7, pp. 1–10 (1998)
6. Ben-David, A.: What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. New Media Soc. **18**(7), 1103–1119 (2016)
7. Ben-David, A., Huurdeman, H.: Web archive search as research: methodological and theoretical implications. Alex. J. Natl. Int. Libr. Inf. Issues **25**(1–2), 93–111 (2014)
8. Brügger, N.: Website history and the website as an object of study. New Media Soc. **11**(1–2), 115–132 (2009)
9. Brügger, N.: Probing a nation's web sphere: a new approach to web history and a new kind of historical source. In: Proceedings of the 2014 ACM conference on Web science. ACM (2014)
10. Brügger, N., Finnemann, N.O.: The web and digital humanities: theoretical and methodological concerns. J. Broadcast. Electron. Media **57**(1), 66–80 (2013)
11. Brunelle, J.F., Kelly, M., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: Not all mementos are created equal: measuring the impact of missing resources. Int. J. Digit. Libr. **16**(3–4), 283–301 (2015)
12. Burger, W., Burge, M.J., Burge, M.J., Burge, M.J.: Principles of Digital Image Processing. Springer, Berlin (2009)
13. Callahan, E.: Cultural similarities and differences in the design of university web sites. J. Comput. Med. Commun. **11**(1), 239–273 (2005)
14. Costa, M., Gomes, D., Couto, F., Silva, M.: A survey of web archive search architectures. In: Proceedings of the 22nd international conference on World Wide Web companion, pp. 1045–1050. International World Wide Web Conferences Steering Committee (2013)
15. Davison, B.D.: Topical locality in the web. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 272–279. ACM (2000)
16. Dougherty, M., Schneider, S.: Web historiography and the emergence of new archival forms. In: ParK, D., Jankowski, N., Jones, S. (eds.). The long history of new media: technology, historiography, and contextualizing newness, pp. 253–266. Peter Lang Publishing, New York (2011)
17. Engholm, I.: Digital style history: the development of graphic design on the Internet. Digit. Creat. **13**(4), 193–211 (2002)
18. Finlay, R.: Weaving the rainbow: visions of color in world history. J. World Hist. **18**(4), 383–431 (2007)
19. Gomes, D., Silva, M.J.: Characterizing a national community web. ACM Trans. Int. Technol. (TOIT) **5**(3), 508–531 (2005)
20. Gorsky, M.: Sources and resources into the dark domain: the UK web archive as a source for the contemporary history of public health. Soc. Hist. Med. **28**(3), 596–616 (2015)
21. Hale, S.A., Yasseri, T., Cowls, J., Meyer, E.T., Schroeder, R., Margetts, H.: Mapping the UK webspace: fifteen years of British universities on the web. In: Proceedings of the 2014 ACM conference on Web science, pp. 62–70. ACM (2014)
22. Hill, B., Roger, T., Vorhagen, F.W.: Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference formula. ACM Trans. Graph. (TOG) **16**(2), 109–154 (1997)
23. Hochman, N., Manovich, L.: Zooming into an Instagram city: reading the local through social media. First Monday 18(7) (2013)
24. Hockx-Yu, H.: Access and scholarly use of web archives. Alex. J. Natl. Int. Libr. Inf. Issues **25**(1–2), 113–127 (2014)

25. Hu, J., Bagga, A.: Categorizing images in web documents. IEEE Multimed. **11**(1), 22–30 (2004)
26. Huurdeman, H.C., Ben-David, A., Sammar, T.: Sprint methods for web archive research. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 182–190. ACM (2013)
27. Huurdeman, H.C., Kamps, J., Samar, T., de Vries, A.P., Ben-David, A., Rogers, R.A.: Lost but not forgotten: finding pages on the unarchived web. Int. J. Digit. Librar. **16**(3–4), 247–265 (2015)
28. Jackson, M.H., Purcell, D.: Politics and media richness in World Wide Web representations of the former Yugoslavia. Geograph. Rev. **87**(2), 219–239 (1997)
29. Jatowt, A., Kawai, Y., Tanaka, K.: Page history explorer: visualizing and comparing page histories. IEICE Trans. Inf. Syst. **94**(3), 564–577 (2011)
30. Keenan, T.: Looking like flames and falling like stars: Kosovo, 'The First Internet War'. Soc. Ident. **7**(4), 539–550 (2001)
31. Klein, M., Nelson, M.L.: Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. Int. J. Digi. Libr. **14**(1–2), 17–38 (2014)
32. Lin, J., Kraus, K., Punzalan, R.: Supporting "distant reading" for web archives. Proc. Digital, Humanities, pp. 239–241 (2014)
33. Manojlovic, I.: The museum of Yugoslav history. The acquisition of the .yu domain. In: Domanovic, A. (ed.). From Yu to Me, pp. 1–5. Firstsite (2014)
34. Marcus, A., Hamoodi, S.: The impact of culture on the design of Arabic websites. In: International Conference on Internationalization, Design and Global Development, pp. 386–394. Springer, Heidelberg (2009)
35. Mazzucchelli, F.: What remains of Yugoslavia? From the geopolitical space of Yugoslavia to the virtual space of the web Yugosphere. Soc. Sci. Inf. **51**(4), 631–648 (2012)
36. Milligan, I.: Mining the 'Internet Graveyard': rethinking the historians' toolkit. J. Can. Hist. Assoc. Revue de la Société historique du Canada **23**(2), 21–64 (2012)
37. Nguyen, H., Weber, M.S.: Internet archives as a tool for research: decay in large scale archival records. In: 2015 IEEE International Congress on Big Data (BigData Congress), pp. 724–727. IEEE (2015)
38. Niu, J.: An overview of web archiving. D-Lib Mag. **18**(3), 2 (2012)
39. Pelet, J.-É., Papadopoulou, P.: The effect of colors of e-commerce websites on consumer mood, memorization and buying intention. Eur. J. Inf. Syst. **21**(4), 438–467 (2012)
40. Phillips, M.E.: What should we preserve? The question for heritage libraries in a digital world. Libr. Trends **54**(1), 57–71 (2006)
41. Robertson, A.R.: The CIE 1976 color-difference formulae. Color Res. Appl. **2**(1), 7–11 (1977)
42. Schafer, V.: Part of a whole: RENATER, a twenty-year-old network within the Internet. Inform. Cult. **50**(2), 217–235 (2015)
43. Schneider, S.M., Foot, K., Kimpton, M., Jones, G.: Building thematic web collections: challenges and experiences from the September 11 web archive and the election 2002 web archive. Digital Libraries, ECDL, pp. 77–94 (2003)
44. Sharma, G., Bala, R.: Digital Color Imaging Handbook. CRC Press, Boca Raton (2002)
45. Silakari, S., Motwani, M., Maheshwari, M.: Color image clustering using block truncation algorithm. arXiv preprint; arXiv:0910.1849 (2009)
46. Thelwall, M., Vaughan, L.: A fair history of the web? Examining country balance in the Internet Archive. Libr. Inf. Sci. Res. **26**(2), 162–176 (2004)
47. Wang, S.-L., Liew, A.: Information-based color feature representation for image classification. In: IEEE International Conference on Image Processing, 2007. ICIP 2007, vol. 6, pp. 353–356. IEEE (2007)
48. Weikum, G., Ntarmos, N., Spaniol, M., Triantafillou, P., Benczúr, A.A., Kirkpatrick, S., Rigaux, P., Williamson, M.: Longitudinal analytics on web archive data: it's about time! In: CIDR, pp. 199–202 (2011)
49. Weltevrede, E., Helmond, A.: Where do bloggers blog? Platform transitions within the historical Dutch blogosphere. First Monday 17(2) (2012)
50. Wesolkowski, S.: Color image edge detection and segmentation: a comparison of the vector angle and the euclidean distance color similarity measures. Master's thesis, University of Waterloo, Canada (1999)
51. Zhang, J., Hsu, W., Lee, M.L.: An information-driven framework for image mining. In: International Conference on Database and Expert Systems Applications, pp. 232–242. Springer, Heidelberg (2001)