8

National Web Studies
The Case of Iran Online

Richard Rogers, Esther Weltevrede,
Erik Borra, and Sabine Niederer

Introduction: National Web Studies

In 2007, Ricardo Baeza-Yates and colleagues at Yahoo! Research in Barcelona published a review article on "characterizations of national web domains" in which they sketched an emerging field that we would like to call national web studies. Of particular interest in the article is the distinction the authors made between studies in the 1990s on the characteristics of *the web* and those a decade later on *national webs* (Kehoe et al. 1999; Baeza-Yates et al. 2007). The term "national web," we feel, is useful for capturing a historical shift in the study of the Internet, and especially how the web's location-awareness repositions the Internet as an object of study. A national web is one means of summing up the transition of the Internet from "cyberspace," which invokes a placeless space of email and packets, to the web of identifiable national domains (.de, .fr, .gr, and so on) as well as websites whose contents, advertisements, and language are matched to one's location. In other words, we are at once foregrounding a change to the object of study and asking how to study the new dynamics online. How should we approach the web after cyberspace? The notion of the national web, we argue, is also useful beyond the conceptual level. It enables the study of the current conditions of a web space demarcated along national lines, as Baeza-Yates and colleagues pointed out in comparing one national web with another. As we would like to pursue here, it also may be useful for the study of conditions not only of the online but also of the ground. That is to say, national web studies is another example of country profiling.

Here, building upon the web characterization work, we provide an approach to the study of national webs that both engages a series of methodological debates (how to study a national web) and provides an overall rationale for their study (why study a national web). Where the former is concerned, we put forward an approach that is cognizant of the multiplicity of user experiences of the web as well as the concomitant web data collection practices (that users may actively or

A Companion to New Media Dynamics, First Edition. Edited by John Hartley, Jean Burgess, and Axel Bruns.

passively participate in). Search engines and other web information companies such as Alexa routinely collect data from users who search and use their toolbars, for example. Platforms where "crowds share" by posting and by rating are also data-collection vessels and analysis machines. The outcomes of these data-gathering and data-counting exercises are often ranked lists of URLs, recommended to users. When location is added as a variable, the URL lists may be country- or region-specific. The same holds for language; namely, websites may be served that are wholly or in part in a particular language. Thus, in practice one is able to speak of country-specific and/or language-specific webs organized by the data collected and analyzed by engines, platforms, and other online devices. There is a caveat. Users of these devices draw upon their own data, and are recursively provided with a selection of considered URLs. Personalization may influence the country- and language-specific URLs served, however much to date the impact on search engine results appears to have been minimal (Feuz et al. 2011). Consequently, the effects of personalization are not treated here (Pariser 2011).

We give the term "device cultures" to the interaction between user and engine, the data that are collected, how they are analyzed, and ultimately the URL recommendations that result. In the case study in this chapter, we discuss a series of device cultures and the kinds of national webs they organize. We discuss a bloggers', an advertisers', a surfers', a searchers', and a crowd-sourced web, each formed by the online devices and platforms that collect their data and ultimately purport to represent or provide in one manner or another a country-specific and/or language-specific web. Put differently, we make use of web devices that "go local" – that is, devices that not only collect but serve web content territorially (usually nationally) or to a particular language group. We recognize that going local has those two distinct meanings, which in certain cases are reconcilable and in other cases are not. An engine may serve language-specific websites originating from inside the country as well as from outside the country in question. For example, in return for a query, the Bolivian "local domain Google" (Google.com.bo) may just as well serve results from Spain as from Bolivia or Columbia, with both being in Spanish. Thus, when discussing the demise of cyberspace, and the rise of a location-aware web, there is a tension between two new dominant ways of interpreting the object of study: national webs versus language webs. We are sensitive to the tension between the two new manners of approaching the web after cyberspace, and are aware that "the local," which as mentioned above is how Google terms its national domain engines, may refer to either a national web a language web or both.

We also discuss approaches to demarcating a national web, including sampling procedures. We are particularly interested in the fruitfulness of research outcomes from both keeping separate as well as triangulating the various parts of a national web – the bloggers', the advertisers', and so on. Are the URLs that are listed as "top blogs" by blog aggregators similar to the URLs that are listed as interesting by crowd-sourcing platforms? Does the list of URLs with high traffic and available advertising space for speakers of a particular language (e.g., Persian) resemble the list of the most visited websites in a related country in question (e.g., Iran)? We

conclude that keeping the parts of the web and the lists of URLs separate may be beneficial, as a national blogosphere may have characteristics different from those of a national crowd-sourced web.[1]

The overall rationale for studying a national web not only implies a critique of the web as placeless space, and as universalized, but also is a means to develop further analyses of relationships between web metrics and ground indicators. That is, another focus of this study is the consideration of digital methods to understand the significance of national web space. By digital methods we mean algorithms and other counting techniques whose inputs are digital objects, such as links and website response codes, and whose application pertain to, but ultimately move beyond, the study of online culture alone (Rogers 2009). We discuss metrics for analyzing the health of a national web, such as its responsiveness, freshness, and accessibility. We have experimented with such analyses before, seeking to diagnose the condition of Iraq (in 2007, some four years into the Iraq War) by looking at "its web" (Digital Methods Initiative 2007). We found a broken web. Iraqi university websites were down or had their domains poached and parked. Iraqi governmental sites were suffering from neglect, with the exception of the Ministry of Oil (www.oil.gov.iq), which was bilingual and regularly updated. In our brief foray into the state of Iraq via the Iraqi web, we sought to develop a series of metrics for diagnosing the health of a web that were both conceptual and empirical.

Blocked yet Blogging: The Special Case of Iran

The case study we will consider in this chapter is Iran.[2] It is in many respects a special case, not least because the term "national web" itself may be interpreted to mean the separate Internet-like infrastructure that is being built there (Rhoads and Fassihi 2011). It is also a special case for the scale and scope of Internet censorship undertaken by the state, which is coupled with the repression and silencing of voices critical of the regime. In other words, the Iranian web is experienced differently inside Iran from outside Iran, which is of course the case for all countries where state Internet censorship occurs. It is also seemingly authored differently from outside than from inside Iran. As a consequence, many Iranians online, either site visitors or authors, whether inside or outside the country, need to cope with censorship. Inside the country, coping could mean being frustrated by it and waiting for a friend or relative to bring news about a virtual private network (VPN) or another means of getting around blockages. It could mean routinely circumventing censorship through VPNs, proxies, Google Reader, and other means. Both inside and outside the country, coping may mean actively learning about (and consciously not using) banned words, and perhaps employing code words and misspellings instead. It could mean self-censorship. Dealing with online thuggery is another matter. For example, one may be warned or pursued by the Iranian cyber army (Deibert and Rohozinski 2010). One copes, or protects oneself, through the careful selection of one piece of software or platform over another, based on which one provides safeguards and

forms of anonymity. One may use wordpress.com for the ease with which one may choose a new email address as a login, or Friendfeed for the capacity to change usernames.

Having mentioned some of the reasons why it is a special case, we also would like to point out that certain general metrics such as site responsiveness and freshness may be put to good use when studying countries such as Iran. For example, if sites are blocked by the state yet still responding and updated, one may have indications of a reading audience, both outside and inside Iran. One may have indications of widespread censorship circumvention, as we have found. Here in particular the retention of the separate webs in our sampling procedure is beneficial. That is, the Iranian blogosphere, or the Iranian bloggers read through Google Reader and indexed by Likekhor, are roundly blocked by the state yet remain blogging. ''Blocked yet blogging'' may be the catchphrase for at least certain vital parts of the Iranian web.

Though perhaps not often recognized as such, national webs are nevertheless routinely created. It may be said that national webs come into being through the advent of geo-location technology, whereby national (or language) versions of web applications (such as Google) are served nationally (e.g., Google.gr for Greece) together with advertisements targeted to locals and information that complies with national laws (Goldsmith and Wu 2006; Schmidt 2009).

There is, of course, further literature to draw upon when studying national webs, from the pioneering ethnographic study of the national web of Trinidad and Tobago (where not global but rather Trini culture is performed) to well-known works on media as a means of organizing national sentiment and community more generally (Higson 1989; Anderson 1991; Miller and Slater 2000; Ginsburg et al. 2002). In policy studies, too, national webs, or portions of them, are increasingly ''mapped'' to inform debates about the extent to which the web, and especially the blogosphere, organizes voice (Kelly and Etling 2008; Etling et al. 2010). Of interest is the related work that seeks to build tools to circumvent censorship so that voice is still heard (Glanz and Markoff 2011; Roberts et al. 2011). In library science, national webs are routinely constructed by national libraries and other national archiving projects, which also have considered how to define such a web (Arvidson and Lettenström 1998; Arms et al. 2001; Abiteboul et al. 2002; Koerbin 2004). There are variously sized national web archives. Countries that have legal deposit legislation not only for books but also for web content (such as Denmark) tend to have notably larger web archives than the countries that do not (such as the Netherlands) (PADI n.d.; Lasfargues et al. 2008).

Defining National Websites, and the Implications for National Web Capture

Archivists' definitions of national webs and national websites are of special interest in our undertaking. How do national libraries define national webs and websites?

What may we learn from their definitional work? For example, the Royal Library of the Netherlands, following similar definitions of a national website from archiving projects in other European countries, defines a website as "Dutch" if it meets one or more of the following tests:

1. Dutch language, and registered in the Netherlands;
2. Any language, and registered in the Netherlands;
3. Dutch language, registered outside the Netherlands; or
4. Any language, registered outside the Netherlands, with subject matter related to the Netherlands. (Weltevrede 2009)

This scheme for what constitutes a national website, or at least one deemed relevant for a national archiving context, has consequences for their collection. Here in the first instance we would like to discuss how a definition affects the collection technique, whether automated or by hand. If one were to begin with sites from the national domain (.nl), those sites (whether in Dutch or other languages) could be automatically detected with software, and in the collection procedure one could remove from the list .be sites (from Belgium, where Dutch, or Flemish, is also spoken) unless they treated Dutch subject matters. (Dutch national web archive users likely would be surprised to come upon Belgian websites stored in it for whatever reason!) This approach by the Royal Library could be described, however, as an editorial approach; websites related to Dutch subject matters and websites in Dutch but registered outside the Netherlands (outside .nl) pose particular challenges to automation, especially when working on a large scale. As a research practice, one would not be able to automate the detection and capturing of these kinds of sites; one would more likely create a list of them before routinely capturing them over time. In the national web characterization studies reviewed by Baeza-Yates et al. (2007) and discussed at the outset, the national domain (known as the country code top-level domain, or ccTLD) is the organizing entity. In practice, however, many countries (or nationals) use URLs outside their national domains, such as .com, .net, and .org. As we note below, for Iran, sites with the .ir ccTLD in fact may not be the preferred starting points for demarcating a national Iranian web. As a case in point, in our data the percentage of .ir sites that is blocked is very low, compared to .coms, for example. Thus, .ir sites seem to have characteristics that differ from other sites authored and/or read by Iranians.

In order to describe the considerations an analyst may have when beginning to demarcate a national web, and at the same time to direct these thoughts specifically toward Iran, we first surveyed a selection of Iranian bloggers about the "Iranian web," and particularly the very ideas of an Iranian website and a national web.

We are particularly interested in contrasting definitions of a national web that are "principled" with those based on device cultures. By principled we mean a priori definitions of what constitutes a national web and a national website, such as that of the Royal Library above. By device cultures we mean the webs that are formed

by collecting and analyzing user data, and outputting leading sites of a country and/or language. We mentioned above some of the consequences of demarcating a national web when national websites of interest to archiving are based on formalist properties of their content: it becomes difficult to make a collection at any scale.

In preliminary research about the notion of an Iranian web, a small survey, undertaken by a New Media MA student at the University of Amsterdam, was conducted of 141 Iranian bloggers using Google Reader (Gooder) in the student's Gooder network (Zarrinbakhsh 2011). A variety of definitions of a national web were put forward, and the respondents were asked to choose which definition was best suited (they could choose multiple answers). From the beginning, the question was met with suspicion, as the term itself was seen as a possible ruse by the Iranian government to create its own Internet, and further isolate the country and the people, as the student reported. In comments on the question, it was written that the Internet is a "free sphere" and ideas of a national web would "limit" such freedom.

The survey asked "What is an 'Iranian website'?" and gave the following possible options:

1. Only in the Persian language
2. In Persian and other languages (and dialects) spoken in Iran
3. Authored by Iranians
4. Related to Iranian issues
5. Accessed by Iranians
6. National domain (.ir)
7. Returned by Google

Note first the expansion of what could constitute a national web beyond what we have related so far, both in the national domain characterization studies and in the case of the constitution of the Dutch web by the Royal Library. In particular, the options "accessed by Iranians" and "returned by Google" are newly added candidate constructs of an Iranian web, where the former treats the Iranian web like a traditional media consumption survey about the sites most visited. The latter option, about Google's relationship with the Iranian web, is more ambiguous. Google could be equated with the web generally, as its entry point. Or, one could find the Iranian web with Google.

As a whole, 12 percent of respondents believed that only Persian websites could be considered national websites; 31 percent checked the box for Persian websites and other languages and dialects spoken in Iran; 45 percent thought that Iranian-produced the content could be counted as part of a national web; 29 percent were of the opinion that everything related to Iranian issues is in the area of the Iranian national web; and 19 percent indicated that websites accessed by Iranians should be considered their national web. It should be noted that some people were very much opposed to this fifth choice; they mentioned that every website can be accessed by anyone, so this item seems to be ill conceived. Only 4 percent of the

total respondents chose websites with the Iranian domain (.ir), which implies that defining national web studies based on the domain only would prove unpopular, and 9 percent thought that websites that are shown in Google search results make up the (Iranian national) web.

Finally, in a follow-up question addressing the issue of any difference between writing from inside and outside the country, approximately one-third of the respondents seemed to agree with the following statement by the communications scholar Gholam Khiabany:

> If Iranian blogs are defined in terms of language, this means omission of a large number of Iranian bloggers who write in other languages, most notably English, while including a number of bloggers from Afghanistan or Tajikistan who write in Persian. Focusing on Iranian bloggers writing inside the country also leads to excluding a large number of Iranian bloggers writing in Persian outside Iran. (Khiabany and Sreberny 2007: 565)

On the basis of these survey findings, and extending Khiabany's argument, the analyst concluded that a national web could be defined as one that is authored by Iranians, no matter their location or language in which they write, and no matter the subject matter. In all, the definition of the national web appears to include sites with content authored by Iranians outside Iran in languages other than Persian, and on issues that may not be related to Iranian affairs. This definition makes it nearly impossible to demarcate an Iranian web! But, in any case, detecting sites authored by Iranians outside Iran in languages other than Persian would require manual work. It may be worth noting that the definition adhered to by the Royal Library of the Netherlands also required manual work, but did not expand its definition of Dutch sites to sites authored by Dutch people abroad in languages other than Dutch, unless the subject matter was Dutch-related.

Having considered and discussed what we have termed principled approaches to defining national websites and webs, we instead chose to analyze the outputs of devices, which we will come to again in more detail shortly. That is, methodologically, we did not begin with a priori definitions of what constitutes an Iranian website, or the Iranian web, however fascinating such definitions may be in a formalistic and ontological sense. Rather, as we will explain and eventually defend, we relied upon the URL recommendations made by dominant web devices and platforms that, through various algorithms and logics, were deemed relevant for a specific country and/or language.

Our contribution to national web studies informs the literature on national web characterization, as discussed at the start of this chapter, as well as that on policy studies (and political science) about the organization of voice online. It also contributes to media theory and web studies by putting forward the national web as an object of study. The overall approach is not only conceptual but also empirical, in that we seek properties of national web spaces that are indicators of conditions on the ground. Such properties could relate to how responsive a national web is at any given time, and how accessible. Are responsive sites also fresh (recently

updated)? Are sites that are blocked still responsive and fresh? We are also interested in more than the technical web data sets, and how they may be repurposed for social study. As alluded to already, for Iran in particular the content of websites is carefully monitored by the state; websites may be blocked and website authors may be pursued. In the following, we put forward an approach to demarcating a national web in order to study its current conditions.

Demarcating the Iranian Web: Studying the Outputs of Device Cultures

The purpose of the research was to demarcate a nominal Iranian web and analyze its condition, thereby providing indications of the situation on the ground. By ''nominal web'' we mean one that is predicated on the means by which it is organized by online devices and platforms as well as how it is retrieved, both by the user and by the analyst. Here we have chosen to demarcate an Iranian web through multiple, dominant online approaches for indexing and ordering that ''go local,'' privileging language, location, and audience, broadly speaking. Working in July 2011, we found that the web given by three crowd-sourcing platforms aimed at an Iranian audience differed from the webs yielded by a marketing tool for Persian-language advertisers, a surfer pathway aggregator of users in Iran, and a search engine delivering .ir sites as well as other TLD sites (.coms, .nets, etc.) from the ''region,'' even though each purported in some general or specific sense to provide the Iranian web. Ultimately, we have chosen to write about the Iranian webs in the multiple, and discuss each web's characteristics. We thereby addressed an issue faced by the analyst when deciding where to start collecting URLs, be it in terms of compiling seed URLs to crawl; stringing together keywords and operators to form a query; consulting lists of top blogs by inlink count, top URLs by rating, or top websites by hit count; and so on. For our analysis, we selected the outputs of the well-known aggregators of either Iranian- or Persian-language websites, in a sense not choosing one starting point but retaining them all (or at least a number of significant ones).

We also took decisions with respect to dealing with the idea that a sample of the Iranian web would follow (only) from knowledge of its population. As we discuss, in the national web research area, one may expect that the population of a web is knowable (in terms of the number of websites, and some categorization of their types), and thus one would be able to make a sample from it and of its types. In thinking through such an undertaking, one may port scan the Iranian IP ranges and establish whether IP addresses respond to the standard http and https ports 80, 8080, or 443. The next step would be to count how many web servers are active within a specific IP range, and then roughly estimate the number of domains. Alternatively, one could consider approaching the Iranian Internet authority or Iranian ISPs for their data, or one could crawl a seed list of URLs (or multiple lists) using snowball techniques and subsequently sift the large catch using language-detection

software and/or whois lookups. When one begins to rely on web services that have ceilings or have issues with spammers and scrapers (which is most if not all of them), the challenges working with of (relatively) large amounts of online data become apparent. One is unable to run batch queries without permission from corporate research labs, Internet administrative bodies, and others. Just when it is becoming interesting, the research focus turns to the administrative, legal, and social engineering arenas, bringing it all to a standstill. Simply gaining access and finishing a large collecting and sifting project becomes a great achievements in itself. While we have undertaken one medium-scale scraping and querying exercise for this research project, we have largely avoided the techno-administrative arena we refer to above, instead seeking to make use of what is available to web users. This was a conscious choice in favor of relatively small data.

Further, we would like to make a case for method to demarcate a national web (or "webs") that is sensitive to the variety of ways in which one enters web space by belonging to particular device cultures, which we largely equate with engine and platform operations rather than in ethnographic sense (where an object may, for example, be equated with its less tangible qualities or "spirit"). Generally, we introduce national web demarcation methods that repurpose web devices that not only "go local" but also capture device cultures. In short, we are interested in capturing national device cultures. Repurposing web devices has two methodological advantages. First, popular devices may be viewed as mediating and quantifying specific usage. The devices do so by recursively soliciting user participation in content production and evaluation. They calculate the most relevant websites by aggregating links, clicks, views, and votes, thereby outputting collectively privileged sources. Second, the definition of an Iranian web is outsourced to the big data methodology used by devices to order content, which combines algorithmic techniques with large-scale user participation. Relatively small data sets are obtained from the output of these big data devices. Put differently, the repurposing of web devices is both a strategy for researchers of small data to sample from a big data set as well as a means to have samples that represent specific outlooks on how to organize and order web content, as we explain in our discussion of the privileging of hits, links, location, likes, and other measures by the platforms and devices under study.

In the analyses, we wish to chart language and other formal features that are in each Iranian web. More conceptually, in our particular approach to national web studies, we also would like to discuss which portions of the web are healthy, in the sense of (still) online and active, and which are broken, in the sense of unresponsive. Additionally, we are interested in the extent to which each is censored or filtered by the state, and whether there is a relationship between responsive (and fresh) websites and filtered websites. In order to pursue the question of whether censorship kills content, which we formulated in a previous (and preliminary) project on the Tunisian web (prior to the "Arab Spring" of 2011), we have developed means to chart changes in a special part of the Iranian web over time. We use time-series data from Balatarin (a leading crowd-sourced platform) that we scraped, comparing the

significant URLs voted up around the presidential elections in 2009 with those of the same time period in 2010 and 2011. We run the hosts through proxies in Iran so as to check for indications of blocking. Generally we have found that Balatarin's collection of URLs is particularly susceptible to blocking. Prior to reporting on the longitudinal analyses, we will next describe the indexing and ordering mechanisms of the web platforms and devices relevant to the Iranian space. The data culled from these platforms and engines are employed to characterize the web types on offer.

Device Cultures: How Websites are Valued, and Ranked

The early web was organized by amateur as well as professional link-list makers, who took on the mantle of a librarian or specimen collector and made directories of websites, organized by category. Professional or "pro-am" website categorization by topic remains, in the larger-scale directories such as Yahoo! as well as in smaller-scale collections, though the practice has arguably declined in the face of other methods (described here) that have become more and more settled as dominant approaches for valuing websites (Deuze 2007; Bruns 2008). These approaches toward valuing websites we couch in technical as well as politico-economic terms – as "the hit economy," "the link economy," "the geoweb," "crowd-sourcing," and "the like economy" – that highlight what is counted, by whom, and/or where. Crowd-sourcing is a term coined by the Internet trade press (deriving from the practice of outsourcing but also described as the "worker-bee economy") for a situation in which both the so-called wisdom and also the labor of the crowd pollinates the beneficiary, often a Web 2.0 company or service (Howe 2006; Moulier-Boutang 2008). The other term we employ, geoweb (or "locative web"), has less of the connotation of a particular kind of economy yet embodies the means by which sites are sourced.

The hit economy, exemplified by the hit counter on early websites, ranks sites by the number of hits or impressions, where unique visitors count. For this view we chose DoubleClick Ad Planner by Google (referred to here as Google Ad Planner), which is a service that ranks sites by audience for the purposes of advertisers. While Iran is not among the countries listed (likely owing to a combination of the lack of an .ir local-domain Google as well as the US economic sanctions against Iran), "Persian-speaking" is among the site-type categories in the available audience analytics. Thus, one Iranian web would comprise those sites that reach a Persian-speaking audience, as collected and ranked by Google Ad Planner. Using the options available, 1500 unique hosts for a Persian-speaking audience were collected from Google Ad Planner.

"Link economy" is a term that describes the rise of PageRank and other algorithms that value links (Rogers 2002). It also involves a shift in URL ranking logics away from an advertisers' model (counting hits) to a more bibliographic or scientometric manner of thinking (counting citations or links). The term is used to characterize

Google Web Search, though the other main component of Google's algorithm user click-throughs. Searching Google for .ir sites (including .ir's second-level domains) as well as Iranian sites in generic TLDs in Google's regional search yielded some 3500 hosts.[3]

Alexa, like other companies offering browser toolbars, collects user location data (such as a postal code) upon registration and, once the toolbar is installed, tracks websites visited by the user (see Figure 8.1). It thereby keeps records of the sites most visited by user location, and a list the top 500 sites visited by users in Iran can be obtained from it.

Crowd-sourced sites such as the most well-known (Balatarin) and its emulators (Donbaleh and Sabzlink) require registration before the user may suggest a link, which is then voted upon by other registered users. Those URLs with the most votes rise to the top. For this exercise we collected approximately 1100 hosts from Balatarin, 2850 from Donbaleh, and 2750 from Sabzlink.[4] In the following analyses we grouped Donbaleh and Sabzlink together, for they share the device culture (crowd-sourcing). Together they resulted in 4579 unique hosts. We treated the other platform, Balatarin, separately because of its status as a highly significant Iranian website. Launched in 2006, Balatarin is considered to be the first Web 2.0 site in Persian, and was recognized as one of the most popular Persian websites in

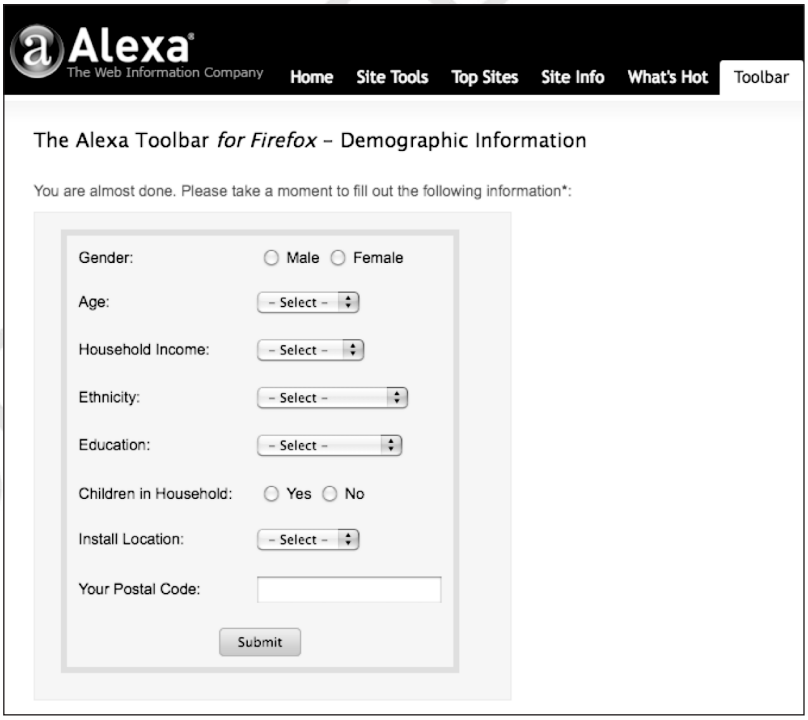

Figure 8.1 Alexa toolbar installation and registration process, with a field for the user's postal code, August 2011. *Source*: © 2012, Alexa Internet (www.alexa.com)

2007 and 2008 (Wikipedia 2011). It was also pivotal for the Green Movement in the opposition before and after the Iranian presidential elections in 2009 (Sabeti 2010).

The introduction of the "like" button and other social counters in social media has brought with it what one may term the "like economy," which values content based on social button activity (Gerlitz and Helmond 2011). Likekhor, as the name suggests, ranks websites by likes; the likes are tallied from Google Reader users who have registered with Likekhor. Google Reader, or Gooder (as some Iranian users call it), is of particular interest because through it one is able to read the contents of websites that are otherwise filtered by the state. Google Reader thus effectively acts as a proxy to filtered websites. At Likekhor the focus is on blogs, pointing up a relationship between Google Reader users and bloggers, or blog readers. From Likekhor we extracted a list of 2600 hosts, collected from a page where all blogs on Likekhor were listed.

Thus, in July 2011 we collected over 10 000 hosts through platforms and devices significant to Iranian users (Google Reader, Google Web Search, and the crowd-sourcing platforms) and two that provide ranked lists of Iranian and Persian-speaking sites (Alexa and Google Ad Planner, respectively) on the basis of data collected from users located in Iran (Alexa) or from Persian-writing users (Google Ad Planner). We will characterize these Iranian webs individually as well as collectively. We have chosen not to triangulate them, for very few websites recur across the them.

Analyzing the Characteristics of the Iranian Webs: Language and Responsiveness

One area of research that we have built on is web characterization studies, in which one of the main difficulties repeatedly discussed is how to obtain a representative sample of a national web or other web types. According to Baeza-Yates and colleagues, the three common types of sampling techniques used in web characterization studies are "complete crawls of a single web site, random samples from the whole web, and large samples from specific communities" (2007: 1). For national webs, which the authors consider to be specific communities, the list comprises websites with the same ccTLD. For many national webs, however, such delimiting would be too partial, certainly for countries where generic TLD usage is prevalent. Our approach seeks to retain the .coms, .orgs, .nets, and so on when deemed relevant for Iranians and Persian-speakers by the devices and platforms upon which we rely.

To the sampling techniques described above, we thus would like to add a fourth type, which could be called "multiple aggregator site scraping" or, more conceptually, device cultures. Google Ad Planner, Alexa, Google Web Search, Likekhor (Google Reader), and the crowd-sourcing platforms (Donbaleh, Sabzlink, and Balatarin) make available either through query results or (dynamically generated) listings websites that are relevant for Iranians and Persian-speakers. In our case, with the exception of the searchers' web (gained through .ir and generic TLD queries

Table 8.1 Percentage of .ir sites in top websites collected from device cultures relevant to Iranians and Persian-speakers, July 2011.

Percentage	*Iranian Web*	*Absolute Numbers*
25	Alexa (geoweb)	126 of 496 hosts
24	Google Ad Planner (advertisers')	370 of 1525 hosts
16	Likekhor (bloggers')	397 of 2541 hosts
12	Donbaleh/Sabzlink (crowd-sourced)	535 of 4579 hosts
11	Balatarin (crowd-sourced)	116 of 1102 hosts

in Google's region search), the percentages of .ir sites among the significant hosts outputted by the devices are relatively low (see Table 8.1). The crowd-sourced web references the fewest .ir sites at just over 10 percent, while the advertisers' web and the geoweb, or web of surfers in Iran, references the highest, at about 25 percent. As noted above, the .ir sites in our overall collection of URLs were much less likely to be blocked than the .com sites. Of the websites that were tested and found to be blocked from inside Iran, 80 percent were .com followed by 6 percent .net and 4 percent .org. The ccTLD .ir had 3 percent of all censored hosts.

Having reviewed how samples are generally made, Baeza-Yates and colleagues compared the 10 national web studies in order to arrive at a core set of measures that are shared across many of them (see Table 8.2). Our characterization of the Iranian web (or webs) has a particular point of departure that benefits from the metrics on offer. In reference to Table 8.2's metrics, in the category of "content" our project shares with Baeza-Yates and colleagues an interest in language, page age, and domain analysis (albeit top-level), and in the category of technology relies on http response codes. The codes yield what we refer to as "responsiveness," which we consider a basic health metric, together with page age (the freshness measure). There are other metrics that we have not employed, though we would like to mention how they could be employed. Whether or not a website is broken

Table 8.2 Metrics commonly used in national web characterization studies according to Baeza-Yates et al. (2007). **Bold** indicates metrics used in this study, but we analyzed the top-level domain over the second-level domain.

Content	*Link*	*Technology*
Language	Degree	URL length
Page size	Ranking	**Http response code**
Page age	Web structure	Media and document formats
Pages per site		Image formats
Sites and pages per domain		Sites that cannot be crawled correctly
Second-level domain		Web server software
		Programming languages for dynamic pages

could be gleaned from link validators, which would refer to broken links on a site. Additionally, establishing whether websites are "parked" or "hacked" may serve to measure abandonment by previous owners. Compared against proxy data, parked or abandoned site analysis may be used to make claims about the effectiveness of censorship, or suppression of voice. Fitness could refer to the "validity" of code, or its correct implementation; Baeza-Yates and colleagues refer to site structure and its "correctness" for a crawler. Other metric types that reside more in the realm of political economy may be of interest in terms of the available ways of expanding the undertaking. For example, media, document, and image formats could give us an indication of the extent to which a national web is proprietary, which from certain perspectives is a health issue.

The Iranian Web and Its Languages

One basic metric seeks to measure the composition of languages in the Iranian web (see Figure 8.2). Persian is of course the official language in Iran (the Unicode system incorporated Persian script in 2001) and it can be detected (Amir-Ebrahimi 2008). For language detection of websites we built a custom tool that makes use of AlchemyAPI and is able to detect Persian as well as the other languages, though not all languages spoken in Iran, as we relate below.[5] The results gleaned using this tool are manually checked.[6] Approximately two-thirds of sites in the Iranian web are in Persian, and English ranks second, making up approximately one in five. Of interest are the proportions of Persian used in the various webs. The results show that the bloggers' space, Likekhor, is on top with 91 percent of its sources in Persian, followed by Alexa's Iran-based surfer's web with 83 percent and the crowd-sourced web with 73 percent. At the bottom are the advertisers' web with 62 percent and Google Web Search with 52 percent. Balatarin, the special case, has

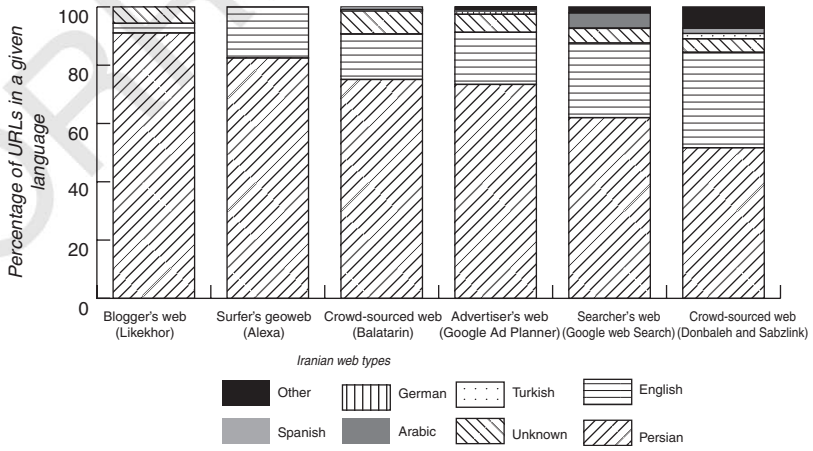

Figure 8.2 The distribution of languages on the Iranian web.

75 percent of its sites in Persian. Thus, there is a significant difference between the webs, including, notably, a Persian-dominant blogosphere (if the Likekhor list may serve as a short-hand reference to such).[7]

Here we can begin to discuss the kinds of webs that could be captured and analyzed if one were to formally define the Iranian web or an Iranian website a priori, as discussed earlier with respect to web archivists' conditions for a national website (in the Dutch example) and survey respondents' ideas concerning a national web (for Iran). The blogosphere and to a slightly lesser extent the geoweb (based on surfers in Iran) are the most closely related to the idea of an Iranian web as Persian-speaking only, though between them they still average over 10 percent non-Persian websites. The Iranian webs with larger percentages of non-Persian sites are the advertisers' and regional webs (from Google's advanced search region option). The advertisers' web is defined as that accessed by Persian speakers as detected by the signals Google compiles on its users and the content it indexes (Google Ad Planner). Both have far higher percentages of non-Persian sites, especially English, though we did not attempt to investigate whether these sites are authored by Iranians or concern Iranian affairs, however that may be defined. There is another, all-inclusive, web one could conceive of a priori that would have implications for the method by which one would construct the object of study. Being all-inclusive in terms of the languages spoken in Iran (Armenian, Assyrian Neo-Aramaic, Arabic, Azeri, Balochi, Gilaki, Kurdish, Lori, Mazandarani, and Turkmen) has consequences for the capturing techniques; of the secondary languages spoken in Iran, the language detection tool employed in this study detects only Arabic, Armenian, and Azeri and not Assyrian Neo-Aramaic, Balochi, Gilaki, Kurdish, Lori, Mazandarani, or Turkmen. To compile such sites one would rely on specialists' link lists, though we did not pursue the matter any further.

The Iranian Web and Responsiveness

To analyze the responsiveness of the Iranian webs, we retrieved the http response status codes with a custom-built tool. The inputs to the tool are the lists of hosts per web that were previously collected. Analyzing the results returned by the response code tool, we found that there are eight commonly returned codes in the Iranian web spaces (see Figure 8.3). The 400 class of status codes indicates that the client has erred in some way. "400 bad request" means that there was an error in the syntax, "403 forbidden" indicates that the server is refusing to respond, and "404 not found" means that the content is no longer available.[8] Commonly returned response codes besides the "200 OK" status are two redirecting response codes: "301 moved permanently" and "302 found." Redirecting is not necessarily an indication of unresponsiveness, and can have a range of reasons, including forwarding multiple domain names to the same location, redirecting short aliases to longer URLs, and moving a site to a new domain.[9] It also may be an indication of a parked website.

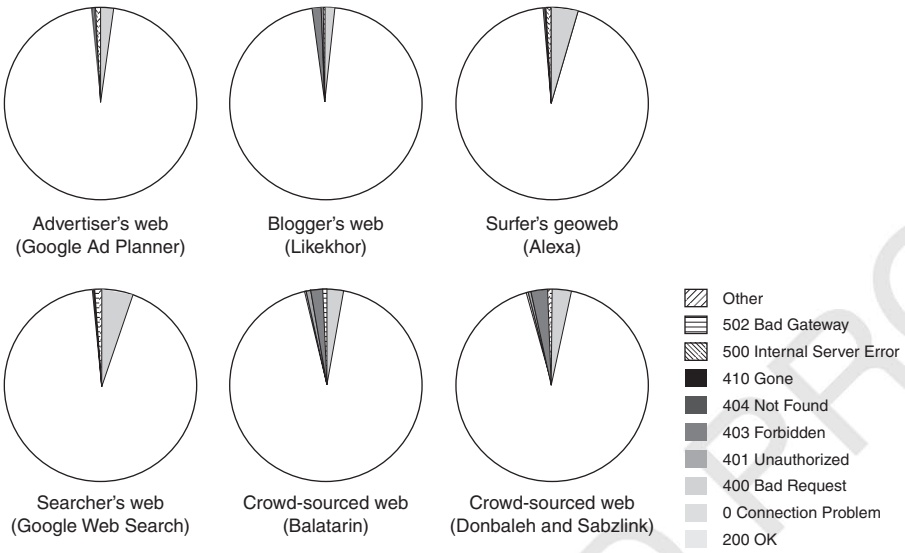

Figure 8.3 The health of the Iranian webs measured by http response codes in the Netherlands.

However, redirects also may be "soft 404" messages to hide broken links (Yossef et al. 2004). In our study, both 301 and 302 were followed if a location header was returned, which mostly resolved in 200 and 404 response codes. "0 connection problem" indicates that the tool was unable to connect to the server; the server may no longer exist, or it may mean that our tool timed out.

The findings of this portion of the study in the first instance indicate that the Iranian webs are relatively healthy overall. Concerning the crowd-sourcing webs (Donbaleh/Sabzlink and Balatarin), 92 and 94 percent of the sites resolved, respectively. The advertisers' space followed by the bloggers' space (delivered by Google Reader users) had the cleanest bills of health, with 96 and 95 percent. Thus, the (Persian-language) advertisers' space and the blogosphere can be considered to be vibrant and healthy.

The Iranian Webs and Internet Censorship

Arguably, web devices are among the most well-informed censorship monitoring instruments. Search engines and platforms receive requests for deleting content – either specific URLs, specific queries, or more general instructions – thereby inviting the creation of an ongoing blacklist as well as a censorship index. For example, it has been reported that to adhere to Chinese government censorship instructions (prior to the redirect to .hk), Google engineers "set up a computer inside China and programmed it to try to access websites outside the country, one

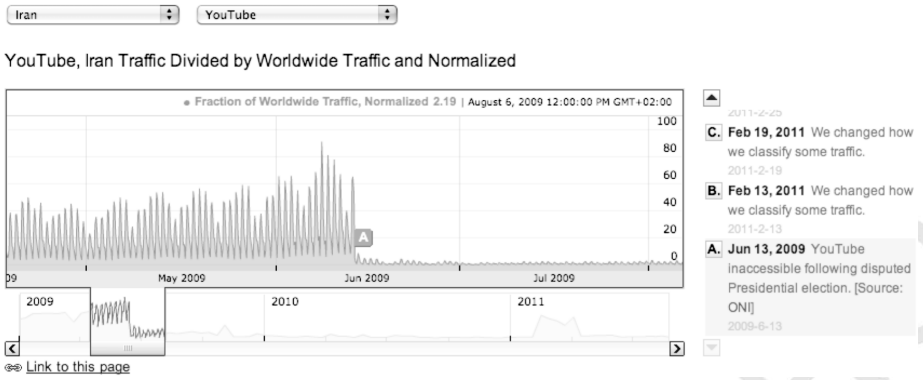

Figure 8.4 Iranian traffic to YouTube comes to a standstill after the 2009 presidential elections. *Source*: Google Transparency Report, August 25, 2011.

after another. If a site was blocked by the firewall, it meant the government regarded it as illicit – so it became part of Google's blacklist'' (Thompson 2006). In the case of the Iranian web, which is among the most aggressively censored webs in the world, there are no reported requests from the government for removal (Open Net Initiative 2009; Google 2011). However, Figure 8.4 shows how Iranian traffic to YouTube increased in the run-up to the presidential elections in June 2009 before coming to an almost complete standstill one day after. The question of interest in this study is to what extent blocking important sites has had implications for the health of the Iranian webs. We therefore checked the collected Iranian webs for availability inside Iran by using proxies. Subsequently, as we will elaborate below, these findings were compared against the basic health measures, responsiveness and freshness. As mentioned above, one of the more remarkable findings is that a large portion of the Iranian blogs is blocked, yet continues to respond and is fresh.

The Censorship Explorer tool, which we have made available at http://tools.digitalmethods.net/beta/proxies, lists (fresh) proxies by country and may be used to check for censored websites. The tool returns website response codes or loads the actual websites in the browser as if you were in the chosen country in question. As a starting point in the censorship research procedure, one often checks website responsiveness in a country that is not known to censor (Iranian) websites (in this case, the Netherlands). Subsequently, one runs lists of hosts through proxies in the country under scrutiny and logs the response codes. Iranian servers typically return the 403 forbidden response code, which is a strong indication of a site being blocked (Noman 2008). Response code checks through proxies may give an indication of specific types of Internet censorship – for example, URL and IP blocking, which includes censorship techniques such as TCP/IP header filtering, TCP/IP content filtering, and http proxy filtering (Murdoch and Anderson 2008). (Other known filtering techniques, including DNS tampering and partial content filtering, are more accurately detected by other means.) Often multiple proxies are used, allowing

the researcher to triangulate proxy results and increase the trustworthiness of the results. For example, "0 connection problem" may be a proxy problem but it may just as well be that the censors is returning an RST package, which resets the connection, effectively dropping it (Villeneuve 2006). Comparing multiple proxies can aid in confirming that it is not a proxy problem. We used 12 proxies, which are hosted in six different cities in Iran and operated by a variety of owners, including Sharif University of Technology and the popular Internet service provider Pars Online. Concern has been voiced that it is "false to consider Internet filtering as an homogeneous phenomenon across a country," considering that both the implementation and user experience of censorship may vary by city, Internet service provider, or even computer (Wright et al. 2011: 5). Taking note of this concern, we selected proxies from different cities and ISPs and subsequently considered the response code returned by the majority.

The results show that approximately 6 percent of the geoweb (29 out of 497) and 16 percent of the advertisers' web (241 out of 1525 hosts) is blocked. The crowd-sourced web has just over 50 percent of the web blocked, with 2411 of 4579 hosts. Balatarin is the most aggressively censored Iranian web space with 58 percent blocked, or 639 of 1102 hosts, followed by the other two crowd-sourcing platforms (Donbaleh and Sabzlink) with more than half of the hosts blocked. Google Reader's web, which in the research work thus far has stood in for the Iranian blogosphere, had 1137 of 2541 sites returning the 403 forbidden code, or 45 percent (see Figure 8.5).

As discussed above, the bloggers' web is largely Persian language, and is one of the most responsive of all the webs under study, with 95 percent of the sites returning "200 OK" response codes. Moreover, it speaks for the Google Reader usage as a vibrant censorship circumvention culture. This study appears to render visible censorship circumvention at a large scale, or at least show that blocked websites are still online. Of the webs checked for filtering, the crowd-sourced sites as well as the Likekhor listing are the most blocked, raising the question not only of the substance of those spaces (we treat Balatarin's below) but also the convenience of the platforms as URL lists for monitoring. While many sites are blocked and still responsive, we are interested in examining those blocked sites for other signs of

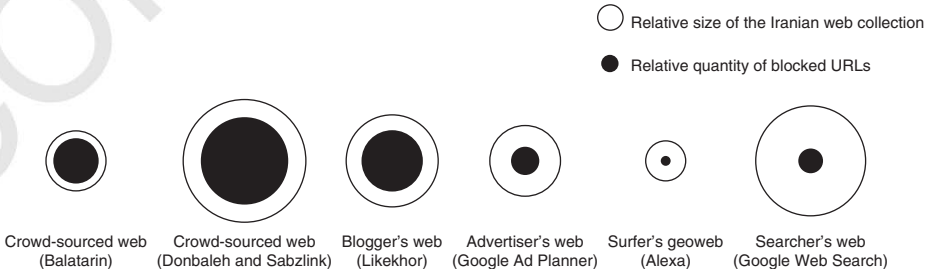

Figure 8.5 Censorship on the Iranian web, August 2011. *Source*: Censorship Explorer tool by the Digital Methods Initiative, Amsterdam.

health. Are they fresh? If the sites are blocked yet responsive and fresh, we have a strong indication of the ineffectiveness of censorship (to date).

The Iranian Webs and Freshness

Having identified the spaces of particular interest to us (the crowd-sourced and bloggers' webs) and found that they were highly responsive as well as heavily blocked, we were interested in pursuing further the question of whether censorship kills content. Or, to put it another way, despite having their sites censored, do the bloggers keep on blogging and does the crowd keep posting and rating? Is there an expectation that the readers can routinely circumvent censorship and thus that content can continue to be recommended, commented on, and so on? Aside from the responsiveness test (which found nearly all of the websites to be online), we wanted to know whether the websites were active. Is the content on the websites fresh? To do this we moved on to study a subset of the webs – the blocked sites in the crowd-sourced and the bloggers' webs. To determine how fresh these sites are, for each host (per list) we asked the Google feed API whether each site had a feed (e.g., RSS or atom). If it did, we parsed the feed with the Python Universal Feed Parser library and extracted the date of the latest post. Overall, 63 percent (5147 of the 8222) of the three webs had feeds. Of the blocked sites in these webs, 71 percent (2986 of the 4189) had a feed. Balatarin had 79 percent blocked sites with a feed (504 of 639 blocked hosts), Donbaleh/Sabzlink had 68 percent (1630 of 2413), and Likekhor had 75 percent (852 of 1137). These were the ones to be checked for freshness.

What constitutes a fresh site? We turned to blog search engines for advice about staleness. In an FAQ about blog quality guidelines, Technorati (2011) states that they "only index 30 days' content, so anything older than that will not appear on Technorati." Similarly, Blogpulse (a search engine and analytics system for blogs) takes 30 days as a measure of fresh content: "A blog's rank is based on a moving average of its citation counts over the past 30 days" (Blogpulse 2011). Thus, freshness is here considered as having at least one post published via a feed in the last month, counted from the moment we last checked for blockage. It is of interest to note that the well-known survey conducted by Technorati in 2008 found that only about seven million of the 133 million blogs it followed had been updated in the past four months. *The New York Times* wrote that the finding implied that "95 percent of blogs [were] essentially abandoned, left to lie fallow on the web, where they become public remnants of a dream – or at least an ambition – unfulfilled" (Quenqua 2009). In stark contrast, we found that 65 percent of our list of sites were fresh. In the crowd-sourcing platform Balatarin, 78 percent of the blocked hosts that had a feed (395 of 504 hosts) were fresh, and in the crowd-sourcing web organized by Donbaleh and Sabzlink 56 percent of the blocked hosts with a feed (915 of 1630 hosts) were fresh. In the Likekhor list, 61 percent (525 hosts) had a latest post date of a month or less when they were tested and found to be blocked. The results confirm

that there is no general indication that censorship kills content on the Iranian web under study. On the contrary, the most severely censored Iranian webs were both responsive and rather fresh.

Conclusion: National Web Health Index

In this study we sought to build upon national web characterization studies and put forward the emerging field of national web studies. First and foremost, we did this by making a methodological plea for capturing and analyzing the diversity of national web spaces, or webs. Rather than predefining national websites, and thereby national webs, according to a principled approach of formal properties (for instance, all websites with ccTLD .ir; all websites in Persian with Iran-related content; or all websites with authors inside Iran), we have concluded that such approaches are often not to be operationalized or automated. Instead, we propose the use of what we term ''device cultures,'' and in particular the Iranian web spaces they provide – bloggers', advertisers', surfers', searchers', and crowd-sourced webs. Device cultures are defined as the interaction between user and engine, the data that are routinely collected, how they are analyzed, and ultimately the URL recommendations that result. We demarcated national webs through devices that ''go local'' – that have location or language added as a value that sifts URLs that are of relevance to Iranians and Persian-speakers. In an examination of the data sets, where we performed TLD analysis of the sample we found that the majority of the collected hosts from the various Iranian webs were .com websites, not .ir; this finding expanded the scope of national domain characterization studies and introduced a method of data collection for a broader national web studies.

Second, both building on and contributing to national web characterization studies, we proposed a rationale: a national web health index. It is conceptualized as a series of metrics, a limited number of which we have employed in this study, most readily responsiveness, page age, and filtering or blockage. (We also performed language detection and TLD analysis.) The contribution of this work to national web characterization studies is twofold. The first contribution is conceptual, in that we propose to repurpose metrics from national web characterization for national web health indices. Are websites responding? Are pages fresh? Are links broken? Is the code valid? Are file formats proprietary? A form of country profiling comes into view. The second is generalizable for countries that face state censorship, and applicable to our case study in question, Iran. We compared the results from the responsiveness tests to those from the filtering tests to investigate whether the blocked sites were still responsive. The approach led us to find a significant number of blogs that were blocked yet still responsive. The finding that there are so many sites that are blocked yet blogging also indicates an audience for the content, both outside Iran and inside, and we believe there to be widespread Internet circumvention in a particular space: the predominantly Persian-language blogosphere authored by Likekhor and Google Reader, which in tandem serve as important filters for Iranian blogs. Although

heavily censored, the Iranian blogosphere as listed by Likekhor remains vibrant. This censored but active space is similar to the crowd-sourced web, organized by Balatarin. Blocked yet posting, Balatarin's recommended websites suggest similar findings to those for the blogosphere: an active audience for blocked websites. In additional, substantive analysis we found that the Balatarin web (as a collection of URLs highly rated and thus rising to the top of the platform) remains clamorous, perhaps even more so than after the presidential elections of June 2009 and the initial rise of the Green Movement. It is a web that neither appears to be widely practicing self-censorship nor one has been cowed and drained of spirit.

Third, we would like to mention certain implications of national web studies as a means for country profiling, as this affects both current and future policies with respect to the web (and its study) and the use of web indicators for social study more generally. As we alluded, regarding our early work on Iraq and the state of its web during the Iraq War in 2007, the study of national web health provides an additional set of measures regarding the current state of universities, ministries, and other institutions. Where is the activity, and where is the neglect? National web studies may also serve as a source of comparative study, and ultimately as a spur to addressing the ill health of one or more webs. Thus, it is an approach to the study of the web that for portions of it could have salutary consequences.

Acknowledgments

We would like to thank the Iran Media Program at the Annenberg School for Communication, University of Pennsylvania, for supporting this work, and Mahmood Enayat for his thoughtful commentary. We also would like to thank the Iranian web culture expert Ebby Sharifi, who gathered with us at the Annenberg School for the Mapping Online Culture workshop in May 2011, as well as Cameran Ashraf, Bronwen Robertson, Leva Zand, and Niaz Zarrinbakhsh, who participated in the 2011 Digital Methods Summer School at the University of Amsterdam.

Notes

1. The data for this study are online at the project website, http://mappingiranonline.digitalmethods.net.
2. According to the International Telecommunication Union, 13 percent of the Iranian population uses the Internet and 21 percent of Iranian households have Internet access (2011). The marketing research reports an urban concentration of users, with "the vast majority (being) young, mostly 15 to 40" years of age (NetBina 2010: 10). Figures on the Iranian diaspora are not available.
3. Google.com's web search was chosen for its dominance in Iran among users of search engines. Data from 2010 list search engine market shares in Iran as follows: Google 90.78 percent, Yahoo! 4.97 percent, Bing 3.64 percent, Ask Jeeves 0.46 percent, and AOL 0.07

percent (MVF Global, 2010). Another marketing research firm lists 2011 market shares in Iran as Google 87.15 percent, Yahoo! 7.27 percent, Bing 4.16 percent, Ask 0.70 percent, AOL 0.12 percent, and Lycos 0.01 percent (Net Applications 2011). According to Alexa, in October 2011, Google.com was the most visited site in Iran, followed by Yahoo.com. We employed site queries in Google.com for the top-level (.ir) as well as the second-level (e.g., .co.ir) domains, and concatenated the results. The query technique did not allow for redirection to a local-domain Google. Because cookies had not been retained, it also did not allow for the personalization of the results.

4. In order to compare the different platforms, we chose to compare hosts instead of full URLs. That is, for Balatarin, we harvested all the URLs listed on the 150 pages of ''hot'' links, resulting in 1102 unique hosts.
5. The language auto-detection tool employed for this work is AlchemyAPI, which for academic researchers allows 30 000 queries per day. AlchemyAPI is at www.alchemyapi .com/api/lang.
6. We manually checked the results that returned sites as English or unknown, and corrected any errors. We have not explored further why dual-language sites are considered to be one particular language by AlchemyAPI. We also would consider using Google as a language detector. The ''unknown'' tags in the cloud indicate that neither the language detection tool nor the researcher was able to determine the language, for in most cases the site was no longer online.
7. Additionally, the Iranian webs show various degrees of language distribution, with Alexa being the least diverse (six languages) and Google Web Search the most (with 36 languages).
8. The http status codes are explained in the dedicated Wikipedia entry: http://en.wikipedia .org/wiki/List_of_HTTP_status_codes.
9. URL redirection is explained in the dedicated Wikipedia entry: http://en.wikipedia.org /wiki/URL_redirection.

References

Abiteboul, S., Cobena, G., Masanes, J., and Sedrati, G. (2002) ''A First Experience in Archiving the French Web.'' Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy (September 16–18).

Amir-Ebrahimi, M. (2008) ''Blogging from Qom, behind Walls and Veils.'' *Comparative Studies of South Asia, Africa and the Middle East*, 28(2), 235–249.

Anderson, B. (1991) *Imagined Communities*. London: Verso.

Arms, W.Y., Adkins, R., Ammen, C., and Hayes, A. (2001) ''Collecting and Preserving the Web: The Minerva Prototype.'' *RLG DigiNews.* 5(2).

Arvidson, A. and Lettenström, F. (1998) ''The Kulturarw Project – The Swedish Royal Web Archive.'' *Electronic Library*, 16(2), 105–108.

Baeza-Yates, R., Castillo, C., and Efthimiadis, E.N. (2007) ''Characterization of National Web Domains.'' *ACM Transactions on Internet Technology*, 7(2), art. 9.

Blogpulse (2011) ''FAQ: How do you Determine Blog Rankings?'' http://web.archive.org/web /20110722023858/http://www.blogpulse.com/about.html.

Bruns, A. (2008) *Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage*. New York: Peter Lang.

Deibert, R. and Rohozinski, R. (2010) "Cyber Wars." *Index on Censorship*, 29(1), 79–90.
Deuze, M. (2007) *Media Work*. Cambridge: Polity.
Digital Methods Initiative (2007) "Diagnosing the Condition of Iraq: The Web View." https://wiki.digitalmethods.net/Dmi/DiagnosingTheConditionOfIraq:TheWebView.
Etling, B., Alexanyan, K., Kelly, J., et al. (2010) "Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization." Berkman Center Research Publication No. 2010–11. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Public_Discourse_in_the_Russian_Blogosphere_2010.pdf.
Feuz, M., Fuller, M., and Stalder, F. (2011) "Personal Web Searching in the Age of Semantic Capitalism: Diagnosing the Mechanisms of Personalization." *First Monday*, 16(2). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3344/2766.
Gerlitz, C. and Helmond, A. (2011) "The Like Economy: The Social Web in Transition." Paper presented at the MIT7 Unstable Platforms conference, Cambridge, MA (May 13–15).
Ginsburg, F., Abu-Lughod, L., and Larkin, B. (2002) "Introduction" in F. Ginsburg, L. Abu-Lughod, and B. Larkin, eds., *Media Worlds: Anthropology on New Terrain*. Berkeley, CA: University of California Press.
Glanz, J. and Markoff, J. (2011) "U.S. Underwrites Internet Detour Around Censors." *New York Times* (June 12). https://www.nytimes.com/2011/06/12/world/12internet.html.
Goldsmith, J. and Wu, T. (2006) *Who Controls the Internet. Illusions of a Borderless World*. Oxford: Oxford University Press.
Google (2011) "Google Transparency Report." www.google.com/transparencyreport.
Higson, A. (1989) "The Concept of National Cinema." *Screen*, 30(4), 36–47.
Howe, J. (2006) "The Rise of Crowdsourcing," *Wired*, 14(6). www.wired.com/wired/archive/14.06/crowds.html.
International Telecommunication Union (2011) *Measuring the Information Society*. Geneva: International Telecommunication Union.
Kehoe, C., Pitkow, J., Sutton, K., et al. (1999) "*GVU's Tenth World Wide Web User Survey*." Atlanta, GA: Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology.
Kelly, J. and Etling, B. (2008) "Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere." Berkman Center Research Publication No. 2008–01. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Kelly&Etling_Mapping_Irans_Online_Public_2008.pdf.
Khiabany, G. and Sreberny, A. (2007) "The Politics of/in Blogging in Iran." *Comparative Studies of South Asia, Africa and the Middle East*, 27(3), 563–579.
Koerbin, P. (2004) "The Pandora Digital Archiving System (PANDAS) and Managing Web Archiving in Australia: A Case Study." Paper presented at the 4th International Web Archiving Workshop, Bath, UK (September 16).
Lasfargues, F., Oury, C., and Wendland, B. (2008) "Legal Deposit of the French Web: Harvesting Strategies for a National Domain." Paper presented at IWAW'08, Aarhus, Denmark (September 18–19). http://iwaw.europarchive.org/08/IWAW2008-Lasfargues.pdf.
Miller, D. and Slater, D. (2000) *The Internet: An Ethnographic Approach*. Oxford: Berg.
Moulier-Boutang, Y. (2008) "Worker Bee Economy." Paper presented at the Society of the Query Conference, Institute of Network Cultures, Amsterdam, The Netherlands (November 13–14).

Murdoch, S. and Anderson, R. (2008) "Tools and Technology of Internet Filtering" in R. Deibert, J. Palfrey, R. Rohozinski, and Zittrain, Z., eds., *Access Denied: The Practice and Policy of Global Internet Filtering*. Cambridge, MA: MIT Press.

MVF Global (2010) "Online Marketing in the top 50 Internet Economies: Lead Generation and Internet Marketing in Iran." www.mvfglobal.com/iran.

Net Applications (2011) "Search Engine Market Share: Iran, Islamic Republic of." https://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4&qpaf=−000%09101%09IR%0D&qptimeframe=Y.

NetBina (2010) "Online Marketing in Iran." http://new.netbina.com/resources/2/docs/Online_marketing_in_Iran_2010.pdf.

Noman, H. (2008) "Tunisian Journalist Sues Government Agency for Blocking Facebook, Claims Damage for the Use of 404 Error Message Instead of 403." *Open Net Initiative* (September 12). http://opennet.net/node/950.

Open Net Initiative (2009) *Internet Filtering in Iran, 2009*. Toronto, ON: University of Toronto. http://opennet.net/sites/opennet.net/files/ONI_Iran_2009.pdf.

PADI (n.d.) "Legal Deposit." *Preserving Access to Digital Information*. https://www.nla.gov.au/padi/topics/67.html.

Pariser, E. (2011) *The Filter Bubble*. New York: Penguin.

Quenqua, D. (2009) "Blogs Falling in an Empty Forest." *New York Times* (June 5). www.nytimes.com/2009/06/07/fashion/07blogs.html.

Rhoads, C. and Fassihi, F. (2011) "Iran Vows to Unplug Internet." *Wall Street Journal* (May 28). http://online.wsj.com/article/SB10001424052748704889404576277391449002016.html.

Roberts, H., Zuckerman, E., and Palfrey J. (2011) "2011 Circumvention Tool Evaluation." Berkman Center for Internet & Society. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/2011_Circumvention_Tool_Evaluation_1.pdf.

Rogers, R. (2002) "Operating Issue Networks on the Web." *Science as Culture*, 11(2), 191–214.

Rogers, R. (2009) *The End of the Virtual: Digital Methods*. Amsterdam: Amsterdam University Press.

Sabeti (2010) "Balatarin: A Battleground for Defining Freedom of Expression." *Iran Media Program*. http://iranmediaresearch.org/en/blog/13/10/11/23/201.

Schmidt, E. (2009) "Prosperity or Peril? The Next Phase of Globalization." Princeton Colloquium on Public and International Affairs. www.youtube.com/watch?v=9nXmDxf7D_g.

Technorati (2008) "State of the Blogosphere 2008." *Technorati*. http://technorati.com/state-of-the-blogosphere.

Technorati (2011) "Blog Quality Guidelines." *Technorati*. http://technorati.com/blog-quality-guidelines-faq.

Thompson, C. (2006) "Google's China Problem (and China's Google Problem)." *New York Times* (April 23). www.nytimes.com/2006/04/23/magazine/23google.html.

Villeneuve, N. (2006) "Testing Through Proxies in China." *Nart Villeneuve* (April 10). www.nartv.org/2006/04/10/testing-through-proxies-in-china.

Weltevrede, E. (2009) *Thinking Nationally with the Web: A Medium-Specific Approach to the National Turn in Web Archiving*. Master thesis. University of Amsterdam.

Wikipedia (2011) "Balatarin." http://en.wikipedia.org/wiki/Balatarin.

Wright, J., de Souza, T., and Brown, I. (2011) "Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics." Paper presented at FOCI'11 (USENIX Security Symposium), San Francisco, CA (August 8). www.usenix.org/events/foci11/tech/final_files/Wright.pdf.

Yossef, Z.B., Broder, A.Z., Kumar, R., and Tomkins, A. (2004) "Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay." *Proceedings of the 13th Conference on World Wide Web*. New York: ACM.

Zarrinbakhsh, N. (2011) "Living as a Criminal: An Ethnographic Study of the Iranian National Web and Internet Censorship." Master thesis. University of Amsterdam.