Niels Brügger; Ditte Laursen; Janne Nielsen

Exploring the domain names of the Danish web

2017

https://doi.org/10.25969/mediarep/12517

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Brügger, Niels; Laursen, Ditte; Nielsen, Janne: Exploring the domain names of the Danish web. In: Niels Brügger, Ralph Schroeder (Hg.): *The Web as History. Using Web Archives to Understand the Past and the Present.* London: UCL Press 2017, S. 62–80. DOI: https://doi.org/10.25969/mediarep/12517.

3

Exploring the domain names of the Danish web

Niels Brügger, Ditte Laursen and Janne Nielsen

Introduction

What does an entire national web domain look like? And how can its development over time be understood? Using the Danish web as our case study, this chapter explores these questions by studying the historical development of the .dk domain names and the .dk domains archived in the Danish national web archive, Netarkivet, as well as in the international US-based web archive Internet Archive. The analysis is a first step in a larger study of the development of the Danish web. This chapter will also address the broad questions above by combining different sources and developing methods that access and analyse materials from the archives in different ways.

An entire national web domain is something that we never experience as such when browsing the web, but nevertheless it is always there as a horizon, as the national context of our browsing. Therefore we need to understand national web domains not only to grasp the national web in its entirety, but also to allow in-depth analyses of web activities within the boundaries of the nation. Large scale analyses of the development of a national web may also be used to shed light on the nation's life outside the web, by comparing outgoing links from the national web domain with migration, immigration, travelling and trade.

Studies of a national web domain inevitably move from the close and detailed reading of individual web elements such as images, web pages or websites to what the literary scholar Franco Moretti calls 'distant reading'. This refers to a reading that zooms out from the individual document to encompass a vast amount of texts (Moretti, 2000). The aim

of a distant reading is to identify systems, structures, patterns and tendencies that transcend the individual texts, at the expense of complete knowledge about each entity in the mass of texts.

The historical study of an entire national web is a rather new field, and only few articles about national web studies exist. Some of the studies focus on national webs at a given point in time and use material archived by the scholars themselves, in contrast to material in web archives (Rogers et al., 2013; Ben-David, 2014, 2016). Clearly historical studies exist, some of which are based on the archived web, but are limited to studying hyperlink networks (e.g. Hale et al., 2014). One study, based on Yugoslavia (.yu, deleted from the internet in 2010) has investigated how the history of an entire country code top-level domain (ccTLD) can be reconstructed, based on URL-lists and material in the Internet Archive (Ben-David, 2016).[1] Hence, best practice is only slowly emerging. In most cases theories as well as methods and the selected source material have to be developed as the research progresses.

This is also the case with the study described in this chapter. It is part of a larger research project 'Probing a nation's web sphere – the historical development of the Danish web'.[2] The aim of the project is to analyse the development of the Danish national web from 2005 to 2015 as it has been archived in the Danish national web archive, Netarkivet. As part of the project we are developing methods and tools to delimit what constitutes 'the national web' at a given point in time. This is necessary because Netarkivet holds several versions of the same online web entity, even within a limited time span. It is therefore imperative to create a smaller collection from the entire web archive, in other words: a corpus.[3] Once the corpus is in place we will perform detailed analyses of the following five focal points: (1) size (size of the entire web domain, of file types and of websites), (2) space (geographical distribution of websites), (3) structure (networks of hyperlinks), (4) aliveness (new/disappeared domain names and frequency of updating), and (5) content (file and software types, language, and semantics, e.g. word frequencies, sentiment analysis/topic modelling). Due to technical limitations it has not yet been possible to perform these planned analyses.

However, the technical challenges have highlighted another way to approach the development of the national Danish web, namely to study the development of the domain names which constitute the Danish web. Lists of all the registered Danish domain names, year by year, can be found in Netarkivet as they were used as the so-called seed-list which was loaded into the web crawler to tell it what to archive. Later in the project we will be studying the archived content itself. First,

we will analyse the archive's meta-content via the list of registered domain names.

Therefore, the aim of the chapter is threefold: first, we investigate how the list of domain names can be studied as a historical source in its own right. Second, we present the results of what the domain names can tell us about the development of the Danish web, and compare the domain name lists to the number of domains that have been archived in Netarkivet and the Internet Archive. Third, we discuss how the results of this study may be used as part of a broader analysis of the development of the Danish web as it was archived by Netarkivet.

Studying the development of a national web domain

When setting out to study the historical development of an entire national web domain, a number of sources may be relevant, from user statistics and texts in news media to oral history accounts, as well as preserved copies of the web of the past. An example of a research project based on a great variety of sources is the French 'Web90 – Patrimoine, Mémoires et Histoire du Web dans les années 1990' (http://web90.hypotheses.org) which studies the development of the web in France in the 1990s. In contrast to a wide range of sources, Hale et al.'s (2014) longitudinal analysis of the UK national web domain, .uk, was based exclusively on the archived web, that is, the web as archived by the Internet Archive. However, one highly relevant source is often overlooked, namely the domain names allocated to a given nation. In order to get an impression of the size and development of a national web domain, access to comprehensive lists of all existing domain names from different points in time would be extremely valuable.

Domain names as a historical source

An inherent question in any study of a nation's web domain is where the national web starts and ends on the global web. The simple answer is that the national web is any web activity related to the nation state in question. However, operationalizing this answer is both easy and a challenge. It is easy since the web comes with its own institutionalized national delimitations, namely the system of ccTLD domain names such as .uk, .dk, .fr. It is fair to say that whatever activity takes place on a ccTLD is related to the nation state in question, thus forming a national domain name space that we can define as 'the national web'. But this

approach can also present a challenge. On the one hand, there may very well exist web material related to a given nation state outside of the ccTLD on other ccTLDs or on generic top-level domains (gTLD) such as .com, .org or .net. Identifying this material can be very time consuming, if it is possible at all.[4] On the other hand, not all nation states can be identified exhaustively by a ccTLD, most notably the USA. There is a national ccTLD, .us, but the vast majority of US related material is found on gTLDs.

Nevertheless, the institutionalized national delimitation mirrored in the ccTLD constitutes an appropriate first step in identifying a national web, or as Ben-David (2016) puts it, the domain name system is 'the Internet's most strict authenticator of nation-states'. The official national lists of domain names are managed by a national organization. The management of a ccTLD is delegated by the global domain name registrar ICANN (Internet Corporation for Assigned Names and Numbers), such as Nominet in the UK, DK Hostmaster in Denmark, and AFNIC in France. These registrars handle the internet's address system within each of the two-letter suffixes for countries and territories such as .uk, .dk, or .fr. Since the lists of ccTLD domain names provided by these organizations constitute a comprehensive inventory of all the web domains within the national domain, it is relevant to include them in any study of the development of a national web. On the one hand, because they delineate the outer limits of the national domain name space and, on the other, because they attest to the development of the national web domain. The domain name list itself can help to answer research questions regarding, for instance, the number of domain names per year, the number (and names) of domain names that have disappeared or been added since last year, and the number of domain names per domain name owner.

Inventories of the physical space and its inhabitants have been known and used as historical sources for centuries: maps, registers of land and real estate, and population registers. However, the historical use of registrars of digital real estate is still uncharted territory. To the best of our knowledge, only one study exists which aims to map a national web domain based on a study of domain names, namely the above mentioned study of the history of former Yugoslavia's web domain .yu (Ben-David, 2016).

This chapter will investigate how the domain names of the Danish ccTLD .dk can be used as a source, and what they can tell us about the development of the Danish web. The principal focus is on 2005–2015, but the study will also look back to the period after 1987 when the Danish

ccTLD was initially registered. The main source is the complete list of domain names from one date each year, supplemented with information about the domain names from other sources, particularly yearly statistical overviews as well as information from Netarkivet and the Internet Archive. In general, domain name lists are not publicly available, but the national registrar DK Hostmaster provides the Danish list to Netarkivet, where it is the basis for the web archive's broad crawls of the entire .dk domain (cf. below). We have had access to the domain name lists for the present study, but they are protected by national privacy acts and must therefore be processed accordingly. This study is therefore in contrast to Ben-David's (2016) study, which deliberately analysed a disappeared ccTLD, .yu, with a view to demonstrating the challenges of reconstructing a domain name list of a disappeared web domain. The present analysis has access to a complete list of domain names for the Danish ccTLD (at least for the period 2005–2015), and it can rely on a national web archive where the web domains to which the domain names refer can be found.

The national Danish web archive Netarkivet and the Danish ccTLD list

The Danish web is preserved in Netarkivet. Netarkivet was established in 2005 by collaboration between the two national libraries – the State and University Library, and the Royal Library. Since then it has collected and preserved the Danish web based on a legal deposit law (Andersen, 2006; Schostag and Fønss-Jørgensen, 2012). Netarkivet is not delimited to material on the ccTLD .dk. The archive also collects material on any other domain name if it is aimed at a Danish audience or treats themes of relevance for a Danish readership (this material is called 'Danica').

Netarkivet uses three archiving strategies: (1) broad crawls where the entire .dk domain and Danica are archived (four times per year from 2012, fewer in 2005–2011); (2) selective crawls where up till 100 frequently updated websites are archived (e.g. news sites on a daily/weekly basis); and (3) event harvests where websites in relation to events are collected (e.g. elections, disasters, sports events, 3–4 events per year). In November 2015 Netarkivet's collection was approximately 654 TB, according to Netarkivet's website (Netarkivet, 2015). A broad crawl in Denmark is a snapshot of all .dk domains as well as Danish websites published under other extensions, such as .com, .org, etc. The broad crawl is performed by harvesting software, which downloads as much web content as possible from the websites on the domain list, including links and

the websites that the domains link to (for more details, see Andersen, 2006). A broad crawl takes two to four months to perform. In the following we will analyse the development of the Danish web based on the lists from 2006, 2009, 2012 and 2015. From 2012, the lists also contain the names of domain name owners. Table 3.1 shows the broad crawls that are studied in the project.

Table 3.1 Selection of broad crawls

Name of harvest definition	Start date	End date
2005–4–10MB (step 1)	16/12/05	10/02/06
2005–4–500MB (step 2)	20/02/06	30/05/06
2009–1–10MB (step 1)	26/02/09	06/03/09
2009–1–4GB (step 2)	10/04/09	06/07/09
2012–1–10MB (step 1)	23/02/12	13/03/12
2012–1–8GB (step 2)	16/03/12	18/04/12
2015–1–10MB (step 1)	22/01/15	28/01/15
2015–1–10GB (step 2)	04/02/15	24/03/15

As can be seen in Table 3.1, the broad crawls are done in two steps. First, all domains are harvested up to a limit of 10 MB (cf. the names of harvest definitions). Most Danish websites contain less than 10 MB, so this step will harvest approximately 85% of the websites (Schostag and Fønss-Jørgensen, 2012). The second step harvests the larger websites, and as Table 3.1 shows, the limit per domain in the second step has been raised over time as the size of the largest websites has increased. The start and end date of the broad crawl and the time spans vary due to different technical issues (Schostag and Fønss-Jørgensen, 2012).[5]

The development of the domain names of the Danish web

The registry of .dk domains is simply a long list of domain names. The list of domain names constitutes a complete inventory of all the domain names on the national ccTLD at a given point in time. Therefore, it can be used to describe the development of the Danish web without looking in the web archive. Since its beginning, Netarkivet has received lists on a recurring schedule from the national domain name registrar. The data are in fixed width format with domain name, registrant name and email information.

Solmark.dk
Solmarken.dk

DOMAIN

Solmarksvej.dk
Solmaster.dk

Figure 3.1 Extract from the .dk domain name list

When handling data spanning ten years, it becomes apparent that no processing and analysing can be performed without standardizing and cleaning up the data.[6] For instance, the data were standardized into UTF–8 because years ago other character encodings were used. Also, the data were cleaned to remove traces from earlier attempts at handling the data. Dirty data were erased; for instance in one year the list had the remains of some sort of pagination headers, so three rows were deleted in 97 instances (one empty, two purple ones, in Figure 3.1). In other years invisible tab characters were detected that could hinder the data load process.

After cleaning, the data were put into the R system for analysis and charting/visualizations.[7] In R, the individual lists were joined into one data frame that became the base for the analysis.

Number of Danish domain names and ownership 2005–2015

In the analysis of the lists, the following questions were asked: (1) What are the total number of domain names over time? (2) How many domain names have disappeared and have been registered compared to previous years? (3) How many domain names have changed hands compared to previous years? (4) What is the relationship between ownership and domains over time?[8]

The number of domain names is, of course, the simplest question.

Figure 3.2 shows that over ten years, the number of domains has been increasing – but also that the increase is decelerating. This could indicate that the number of domain names is stabilizing. This result is in fact in line with a similar result in a study of the .uk domain (Nominet, 2013). Of course, the result says nothing about the number of active and non-active domains.

For the second question – how many domain names have disappeared and have been registered compared to previous years? – the chart illustrated in Figure 3.3 was created.

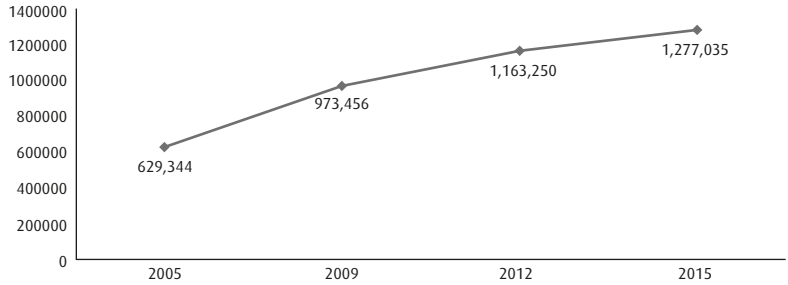

Figure 3.2 Number of .dk domains over time

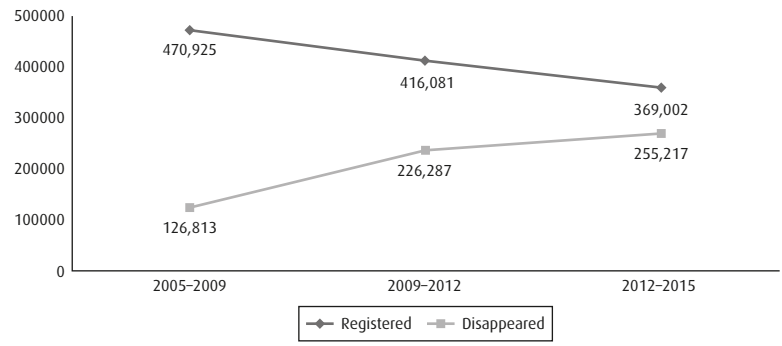

Figure 3.3 Registered and disappearing .dk domain names over time

Figure 3.3 shows that from 2005 to 2009, 126,818 domain names disappeared, and 470,925 domain names were registered. From 2009 to 2012, 226,287 domain names disappeared, and 416,081 domain names were registered. And from 2012 to 2015, 255,217 domain names disappeared while 369,002 domain names were registered. Thus, there has been an increase in the number of disappeared domain names over a three-year interval and a decrease in the number of registered domain names in the same three-year interval. That the two lines approach each other correlates with the gradually slower increase in the total number of domain names indicated in Figure 3.2. As the two lines get closer, the line indicating the total number of domain names will approach the horizontal. Interestingly, Figure 3.3 also says something about the Danish web domain's dynamics or 'aliveness'. At first glance, it looks very dynamic, with many domains being registered and many domains disappearing. However, if we look more closely at the total number of domains that change (that are either registered or disappear), we find

that the numbers add up to approximately 600,000 in all three intervals (2005–2009: 597,738; 2009–2012: 642,368; 2012–2015: 624,219). The dynamics or aliveness of the domain names can therefore be said to be stable over the ten years. This stability can also be seen in the way that the two lines are almost symmetrical around an invisible horizontal line around the number 300,000. In other words, the relationship over time between the increase in disappeared domain names and the decrease in registered domain names is stable.

For the third question concerning the number of domain names that have changed hands over time, we can only compare data from 2012 and 2015, as shown in Table 3.2.

The ratio of domains to owners is approximately the same in 2012 and 2015, with an average of around 2.3 websites per owner. When studying this relationship, however, we find that looking at the average might not be the most relevant way to approach the numbers, as in reality, the domains are not evenly dispersed. In both 2012 and in 2015, just short of 10% of the total number of owners owned 50% of the Danish domains. In addition, in both 2012 and in 2015, if an owner owned more than three domains, s/he belonged to the top 10% of domain owners.[9] When analysing the changes in domain name ownership to answer our third question, we find that in 2015, 14% of the domains from 2012 had changed owner.

In relation to the fourth question – what is the relationship between ownership and domains over time? – the chart in Figure 3.4 shows the results for 2012.

There is no visual difference between 2012 and 2015, and hence no change over the three years. Notably, however, there are two owners who own more than 3,000 domain names, while most owners own one or two domain names.

All four questions are simple questions which reveal something about the development of the Danish web over ten years. The results can be investigated further by means of qualitative analysis. For instance, a closer look at the (types of) domains that have disappeared could uncover interesting patterns. Aspects like these will be studied at a later point in the project.

Table 3.2 Number of .dk domains and .dk owners

Year	Domains	Owners	Anonymous
2012	1,163,250	513,326	46,727
2015	1,277,035	549,978	58,710

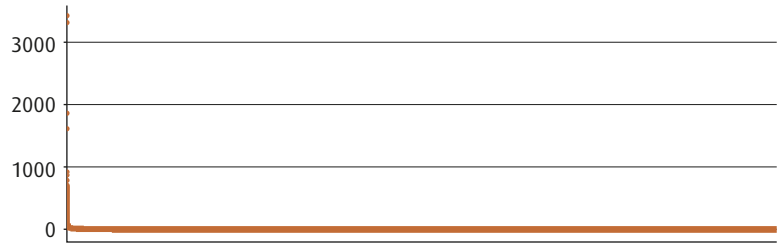

Figure 3.4 Relationship in 2012 between ownership and domains (anonymous registrants removed)

The above results can be further elaborated and put into perspective when combined with data from other sources containing information about national Danish domain names. We have done this in two ways. First, we expanded our analysis of the number of domains over 2005–2015 with data from other sources for the period 1987–2005. Second, we compared our results with data on the number of Danish domains for the period 2005–2015 in Netarkivet and in the Internet Archive, respectively, to see how many of the available domains have actually been archived.

Danish domain names before 2005

In 1987, the internet domain .dk was created. According to an early issue of the magazine of the Danish UNIX User Group *DKUUG-nyt* no 18 (Storm, 1988), the number of registered domain names grew from 49 in 1987 to 70 in 1988. For the years 1989–1995, it has not been possible to locate information on the number of registered .dk domains. But for 1996–2004, a statistics web page from the Danish ccTLD registrar DK Hostmaster's website was found at the Internet Archive.[10] By interpolating from 1988–1996, the chart from 2005–2015 can be expanded as shown in Figure 3.5 (Laursen and Møldrup-Dalum, 2017).

Figure 3.5 shows a slow increase in the years 1987–1997, a steady increase from 1997, a steep increase taking off in the late 1990s, and a slower increase from 2010. There may be various reasons for this development, among which the following three are plausible and could be borne in mind. First, since domain name owners probably prefer as short a domain name as possible, the number of potential names will gradually diminish over the years. Second, the increase in registered domain names correlates with the spread of internet use in Denmark during the

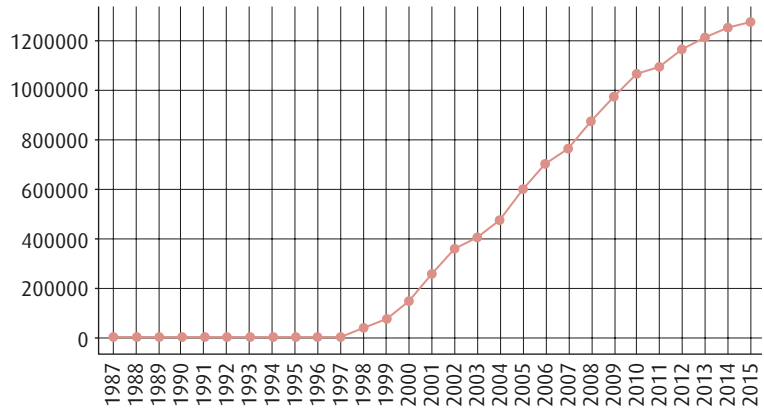

Figure 3.5 Number of .dk domains over time

same period, but with a delay of 2–3 years. The number of internet users slowly increased until 1996 (5%), followed by a steep increase which ended in approximately 2006 (87%) when the curve flattens out until internet access reached 96% in 2014 (Millennium Development Goals Indicators). Not surprisingly, once people have access to the internet, more content is needed, and thus more web domains for content are registered. Third, in 2009 the Danish web domain registrar DK Hostmaster ran a campaign against so-called 'domain name sharks' who bought domain names for 'typosquatting', that is domain names that were misspellings of frequently used domain names (Berlingske Business, 2009).

The Danish domain names in Netarkivet and in the Internet Archive

Our analysis of the domain name lists was compared with data from the archives showing which Danish domains have actually been crawled and archived in the period 2005–2015 to see whether the domain name lists match what is found in the archive. A comparison between the .dk registry list and the domains archived in Netarkivet is shown in Figure 3.6.

As Figure 3.6 shows, more .dk domains are found in the crawled data than on the domain name registry list. This can be explained by differences in time: the registry list is a moment in time, while the crawled data covers a period of time. As time passes, new domains are born. Thus, the two datasets offer two fundamentally different views on the Danish web, where one is no more correct than the other. In addition, Figure 3.6 indicates that the difference in numbers between the registry list and

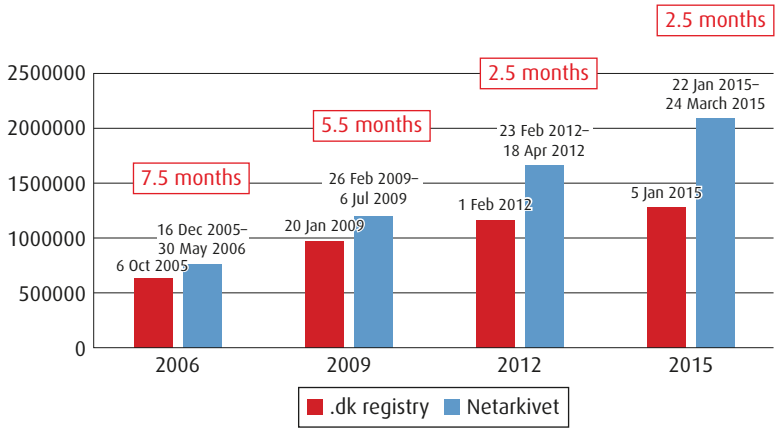

Figure 3.6 Number of domains in the .dk registry list and in Netarkivet

the crawled data increases over time. This could be a sign of aliveness, that there is an increase in the speed at which domains are registered. However, the data are skewed because the crawled data are cumulative – so all the known domain names in the archive are included, even though some may not be active anymore. What we could have done was to exclude domain names with 0 bytes harvested. However, even if we had done that, the data would still not be directly comparable: not only are we trying to compare one moment in time with a period of time, but we are also working with different time spans of the crawls. This means that a comparison between the two kinds of data (and even between the different crawls) has to be done carefully, and taken into consideration when analysing the results.

If we then compare our results with the data from the Internet Archive, the outcome is as shown in Figure 3.7.[11]

Figure 3.7 shows a lot fewer .dk domains in the crawled Internet Archive data than on the domain name registry list. However, again, the data are not directly comparable since the Internet Archive's data, like the numbers from Netarkivet, are based on crawl logs and the .dk registry is not. In addition, data are not from the exact same periods of time. The dates of the .dk registry precede the dates from the Internet Archive, and also the time spans differ: 7.5 months (2006), 5.5 months (2009), 2.5 months (2012) and 2.5 months (2015). Finally, and most importantly, the Internet Archive time spans may or may not cover the archive's broad crawls. Because the intent was to compare crawl log data from the two archives from the same time span, the Internet Archive's

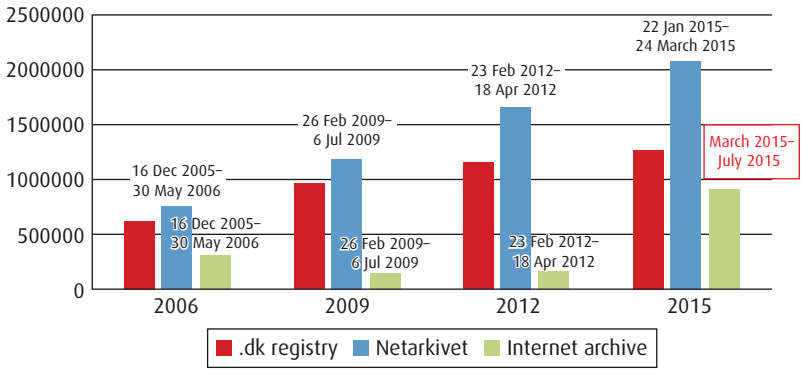

Figure 3.7 Number of .dk domains in the .dk registry, Netarkivet, and the Internet Archive

time periods from 2005, 2009 and 2012 correspond with the start and end date for broad crawls in the Danish web archive. In retrospect, it might have given a more accurate picture had we used the number of .dk domains from broad crawls in the Internet Archive, while still choosing crawls that were as close as possible to the date of the .dk registry list and the dates of Netarkivet's broad crawls. A comparison of broad crawls from both archives would have enabled a less biased result. In 2015, the Internet Archive data do in fact cover an Internet Archive broad crawl, according to information from the archive. Noticeably, this is the year when the number of .dk domains in the Internet Archive is closest to the number of .dk domains on the registry list, that is, 28% less than the .dk registry list.

Comparing numbers of domains, however, does not take into account that domain names may not be the same in the two data sets. For this reason, domain names in the Internet Archive were compared with domain names on the .dk registry list.

Figure 3.8 shows that the Internet Archive contains .dk domain names not found in the .dk registry list, even though the Internet Archive in total contains fewer domain names than the .dk registry list. The difference between domain names in the .dk registry and in the Internet Archive can be explained by the same fact as mentioned above in relation to comparison between the .dk registry list and the domain names in Netarkivet. The new domain names appear in the time span of the crawl (cf. Figure 3.6). For instance, the .dk registry list for 2006 is from 6 October 2005, while the domain names from the Internet Archive are from 16 December 2005 to 30 May 2006. This makes it likely that the

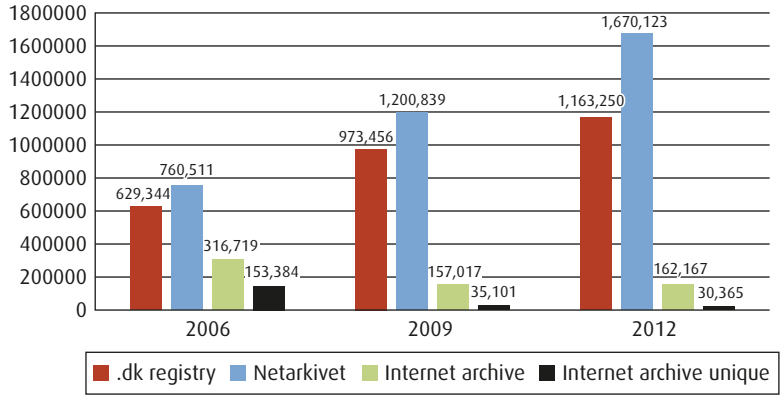

Figure 3.8 Domain names in the Internet Archive not found in the .dk registry

Internet Archive will contain some domain names that do not appear on the .dk registry list.

However, the data offer several possible explanations. One possibility is that the Internet Archive is bad at capturing Danish domains: In 2012, for instance, the Internet Archive collected only about 14% (162,167) of the number of domain names found on the .dk registry list (1,163,250) (cf. Figures 3.2 and 3.7). But from another perspective, the Internet Archive collected about 100% of the statistical increase in newly registered domains from 2009 to 2012 (cf. Figure 3.2). This could be a sign that the Internet Archive is actually very good at capturing new domain names (and that only the new ones are captured). A more likely explanation, however, is changed harvesting settings, which gives bad data or bad calculations. Again, this makes comparing the data a complex matter.

In summary, the number of domains in the Internet Archive does not correspond to the number of domains on the .dk list. The Internet Archive has the aim of capturing all domains and following the links of domains to do so. Consequently, recently registered .dk domains or .dk domains with no or very few ingoing links will have a hard time getting captured. Further studies can provide more insight into the extent to which the difference can be ascribed to the number of .dk domains recently registered as against the number of .dk domains with no ingoing links. Moreover, the Internet Archive captures domains not found in the .dk registry list. This makes it likely that the Internet Archive complements the Danish web archive with regard to some domains. Further

studies can specify the relation between domains on the .dk list and the .dk domains in the Danish web archive. In theory, the domains should be the same, but since a broad crawl takes more than two months, domains may have disappeared from the web before they were crawled. Moreover, .dk domains not on the registry list, i.e. domains that have appeared since the list was made, may have been captured if other .dk domains linked to them.

Finally, a comparison of crawl log data from a broad crawl from both archives could provide a more accurate picture of the capturing of the .dk domains in the two archives and the development of this capture over time. However, a complete comparison will probably not be possible if we take into account that periods of crawling differ and that domains are appearing and disappearing. Even an experiment that started the crawls at the same point in time would make the periods of crawling differ, since different settings and different scopes in the two archives would make crawling end at different times. For this reason, different archives will always complement each other to some extent.

Domain names and archived web

The study of domain names is significant in itself, but it also constitutes an important element in a more comprehensive analysis of the entire ecosystem that constitutes a national web domain. In this section, we will recapitulate what the analysis of the domain names tells us about the Danish web domain, before briefly outlining some of the ways in which a domain name analysis and an analysis of what can be found in a web archive can supplement each other.

What domain names can tell us about the Danish web domain

Digging into the development of domain names can tell us something about three things at least. First, the size of the Danish web can be described through the number of domain names on the Danish ccTLD. However, size understood as the number of domain names does not say anything about how big the Danish web domain is in terms of bytes or number of digital objects. The analysis of the development of the number of domain names indicates that the Danish web name space has grown steadily, until it reached a level where the pace of growth slowed down. To put this another way: the number of street names has increased which in itself is significant, but this does not tell us anything

about how broad or long the streets are, how many houses are on them and how big those houses are.

Second, analysis of the domain names tells us something about the 'aliveness' of the Danish web domain. In the early years, many new domains were registered; the Danish name space was much like a 'wild west' with new streets and houses appearing rapidly. During that time, fewer domains were disappearing compared to later in the study period when the number of disappearing domains began to close in on the number of registered domains. Whether the increasing number of disappearing domains in the later periods signals that more of these early 'streets and houses' are being removed at a more rapid pace, we can only speculate at this point. This could be a subject for further study. Some registered domains might disappear within the same intervals as they were registered, so studies could also be done in order to determine how often this is the case. In addition, it would be interesting to study the average lifespan of a Danish website throughout the period studied.

Third, the domain name lists are very indicative when it comes to ownership. They chart a domain name space where, on the one hand, only a limited number of domain names change hands and, on the other, that this is a space that can be charted as a long tail. Many people each own a very limited number of domain names, and few people own a relatively large proportion of the domain names. This can be likened to a physical space with many small home owners and a few big land owners, or maybe even property speculators.

All three types of result could be correlated with the nation's life outside the web by looking into information found in other sources. For example, by comparing the number of domain names with the number of households with internet access, or by finding more information about the owners who have many domain names.

Domain name studies and archived web content

It would not have been possible to arrive at the results reported here regarding the Danish web domain simply by looking into the web content in the web archive. This is in itself indicative of the value of studying domain name spaces. However, this should not overshadow the fact that combining domain name analysis with analysis of actual archived web content opens up new avenues of inquiry. On the one hand, the archived web can shed light on the development of domain names, and, on the other hand, domain name analysis can help us understand the archived web. Let us have a closer look at these two avenues.

Although the analysis of domain names provides valuable insights, for instance, about which domain names have been established, it does not tell us anything about these web domains: were they actually used by the owner, or were they just an almost empty web page 'under construction'? And what could one actually find on these websites? Following our earlier analogy: what did the streets and houses actually look like? An analysis of archived web content can enhance the analysis of the domain names.

However, the opposite is also the case, since the results of the domain name analysis can supplement analyses of the archived web content and function as a stepping stone to the archived content in at least two ways. The first concerns the completeness of the web archive, the other concerns the generation of new research questions and hypotheses.

The domain name list itself is an important key to evaluating the completeness of how much of a national web domain is in fact a web archive. As an inventory of all the domain names on the national ccTLD, the list can measure the completeness of archived web content in studies where a nation's web domain is delimited by the ccTLD. Access to the historical ccTLD domain name list is particularly important if the corpus of the national web domain is extracted from a web archive which has not been based on a ccTLD domain name list, as is the case with the Internet Archive. If the established corpus is not compared to the domain name list from the same point in the past, however, it is difficult to evaluate the completeness of the corpus at the domain name level. In short, a complete domain name list can help us establish to what extent the archived content actually mirrors 'the nation's web' of the past. However, things may prove to be more complicated than this. A complete domain name list does not in itself provide a solid baseline for what the Danish web actually looked like at a given point in the past, simply because archiving takes time. Each of Netarkivet's broad crawls starts with a comprehensive seed list in the form of the authoritative ccTLD domain name list. However, a comparison of the domain names archived in Netarkivet and those archived in the Internet Archive reveals that there may have been Danish web domains that are not included in the broad crawl based on the ccTLD list. This is simply because web domains are likely to have been registered or to have disappeared during the two to four months it takes to archive the entire Danish web domain. An analysis of domain names can draw our attention to the tension between the static list of ccTLDs that is used at the launch of each broad crawl and the dynamic evolution of the domain name list during the time it takes to archive the ccTLD. There is no easy way to solve this problem. The longer it takes to archive the ccTLD, the greater the chance for an inconsistency between

the initial list of ccTLDs and the evolving list. With a shorter archiving time, the possible inconsistency is smaller, but at the expense of fewer web domains being archived. However, the insights revealed in this study may indicate a need to combine different collections of the archived web.

The second way in which domain name analysis can supplement the archived web content is that it can help generate new research questions and hypotheses. The following two examples can illustrate this. First, one could investigate 'the disappeared web'; that is, all the web domains that have disappeared year by year: What did the disappeared web domains look like? Have specific types or genres of websites disappeared? And are there any patterns or trends in these types compared over time? Second, one could dig deeper into ownership and investigate the many web domains that are with the same owner: Do they belong to a specific content genre, or are they diverse? And how are they interlinked? For instance, a hyperlink analysis of the web domains having the same owner could identify link patterns and maybe tell us something about the extent to which these websites do or do not cluster. In addition, one could correlate the postal address of the owner with actual geo-information on the websites (postal codes, city names, etc.) with a view to investigating whether the web domains 'live' at the same place as the owner. One could investigate the real estate domain of the name landscape. Are there any patterns with regard to content type for the web domains that are often passed to another owner? A final subject for study are the web domains that have 'disappeared' because they were never archived in Netarkivet, having fallen prey to the temporal lapse between the initial ccTLD list and its evolution during the archiving process. Once these web domains have been identified in another web archive such as, for instance, the Internet Archive, one could look for patterns in terms of content or genre.

Conclusion

Many levels of analysis are necessary to derive a comprehensive analysis of the entire ecosystem that constitutes a national web domain. In this chapter, we have taken the first steps in answering the big question(s) of what an entire national web domain could look like and how the concept has developed over time. Our study was based on the historical development of the backbone of the Danish web: the national domain names (ccTLD). This approach has focused on the analysis of historical changes in the .dk domains as they appear in three different sources: lists from the

Danish national registrar, the Danish national web archive, Netarkivet, and the international web archive, Internet Archive. The analysis of the domain name list shows that the number of domain names has increased over the years but that the pace has changed. From a slow start at the end of the 1980s, there was a lot of activity from the late 1990s more or less corresponding with the spread of internet use in Denmark. Since 2010 there has been a tendency for the curve to level off, but it still shows a steady upward slope. This is not surprising. We would expect the Danish web to become gradually larger (here understood as the number of domains) following the general spread and growth of the internet. It has become the norm for companies, institutions and organizations to have their own website, often on their own domain.[12] While the number of domains is increasing, we see that the relationship between registered and disappearing names is relatively stable, highlighting that the dynamics of the Danish web are more complex than just the appearance of more domains. Another aspect relates to the ownership of domains and studies of the top 10% of domain name owners might shed light on parts of the dynamic in the Danish domain name scape.

An important lesson from this study is that the three datasets – the .dk registry list and the list of archived domains from each of the two web archives – offer different insights into the development of the Danish web. Combining them contributes not just to furthering the understanding of each data set, but also to understanding the complete picture of the ecosystem. It is important to supplement the results of the domain name analysis with more analyses of the archived web content, and we propose to do this by creating a corpus for each year and analysing these corpora focusing on the size, space, structure, aliveness and content (as described in the introduction to this chapter). Both approaches constitute valuable methods to the understanding of the evolution of a nation's web domain, and they could both be included as best practice in the toolbox of similar studies in the future. Another way to enhance the results is to combine a quantitative approach with qualitative analyses, studying selected websites in more detail. Hence, a multitude of approaches and sources could possibly be included in further research, and they will all be useful in gaining a comprehensive understanding of the historical development of a national web.