Defining a National Web Sphere over time from the Perspectives of Collection, Technology and Scholarship

By
Eld Zierau, The Royal Library of Denmark, Søren Kierkegaards Plads 1, DK-1016 København K
Niels Brügger, Aarhus University, Helsingforsgade 14, DK-8200 Aarhus N
Jakob Moesgaard, The Royal Library of Denmark, Søren Kierkegaards Plads 1, DK-1016 København K

Abstract

This paper describes a framework supporting definition of how to automatically identify national webpages outside a country's top level domain. The framework aims at a definition that can be put into operation in order to make automatic detection of national web pages. At the same time the framework aims at a definition that can be reused independent of changed behaviours on the net, changes in jurisdiction and changes in technology. A crucial point in this framework is that the perspectives of collection, technology and Scholarship are present in decision making.

The framework origins from a study that aimed at evaluation of different two different strategies for automatic identification of national webpages outside a country's top level domain; one strategy was based on data from Internet Archives wide_005 world wide webcrawl, and the other was based on a local web crawl based on bulk harvests from the Danish national web archive, Netarkivet. However in both cases a definition of national webpages was needed. Thus the creation of the framework was a prerequisite for the rest of this study.

Motivation of the study and framework is based on the fact that human communication activities are moving more and more onto the internet. This means that a lot of present and future research in the 20th century information flow depends on optimised collection and archiving of such information in web archives. Web archives often reside within national cultural heritage institutions, regularly having a collection scope outlined within some form of legal deposit legislation.

The challenge to define "national webpages" showed out to be far from trivial, and in creation of the framework it quickly became obvious that such a definition requires input from three important perspectives in order to make qualified decisions. In this paper this definition is based on input from three important fields represented by each of the authors, representing the perspectives of scholarship, the Danish web Archive, and computer science. This represents the perspectives of collection, technology and scholarship, which are all very different but also crucial perspectives when formulating definition of national webpages that is basis for actual collection and thus consequently form a web archive.

Besides the non-trivial need for "national webpages" definition, the study also found reason for arguing that it is necessary to repeatedly adjust web collection strategies within a web archive. The conditions for web collection are constantly changing. Even over a five year period we see: change in technology that can assist in collection, change in human behaviour moving away from countries top levels domains and out on .com, .org etc., and changes in jurisdiction influencing the way that the web can be collected technology, thus regularly adjustments of what is national web pages may likely be needed. Therefore the presented framework consists of a list of general criteria as basis for adjustment of web collection strategies which can be made operational in a specific context taking into account the three perspectives.

Introduction

Web archiving contents has in the last decade become increasingly important for research, as the human communication activities are moving more and more onto the internet. Thus, a lot of present and future research in the 20th century information flow depends on proper collection of the World Wide Web.

This paper will argue that the lessons learned during the last decade points at the need for repeated adjustment of collection strategies. Optimal web collection strategies need adjustment according to technical evolutions and change behaviour in use of the internet. The paper argues that three perspectives are needed in such evaluations, namely the *perspectives of collection, technology* and *scholarship*. Each of these perspectives is represented by each of the authors.

The motivation for the research presented here has emerged from a need to define a national web sphere, as the basis for a larger research project: the WebDanica project[1] which investigates two strategies for *automatic* collection of web material outside the .dk *ccTLD*, one strategy inspired by the Czech web archive [14] and one strategy using the Internet Archive world wide webcrawl [7].

This paper also sketches a general approach to define and maintain strategies of how to collect relevant web material as automatically as possible over time. The focus is national heritage as the responsibility to provide access to web archives is often defined on a national basis[2], where a nation states risk abandoning parts of the national cultural heritage [2].

Automation is needed in order to overcome the challenge of the size of web collections which is too large to be collected manually. This is also evident from work in the ever growing International Internet Preservation Consortium (IIPC) [6]. New countries continuously join IIPC as consequence of a growing awareness of the importance and challenges in automated collection of the internet. Frequent collection is needed in order to catch web materials that are volatile[3] and therefore quickly disappears [4].

During the last decade there have been made many technical adjustments to harvests of the internet as the technology evolved. For example harvest challenges for Flash. New technical challenges are constantly emerging as the technology evolves.

Collection strategies also need adjustment due to changing internet user behaviour. There are today evidence that there is a continuous growth in size and change in contents of outside-ccTLD [10], e.g. a lot of activities have moved out of country code Top Level Domain (*ccTLD*) (e.g. .uk) and into more international TLD (e.g. a political party's Facebook.com page or country specific like www.slovakia.org). Another example is the emergence of the so-called social media domains such as Facebook, YouTube and Flickr. Such domains will not be part of a nation's *ccTLD*, but contains sub-domains which can be considered.

[1] This project is partly financed by the Danish Ministry of Culture

[2] E.g. IIPC members coming from web archives located at national organisations.

[3] Examples are MySpace (http://www.telegraph.co.uk/technology/10173232/MySpace-users-threaten-to-sue-after-years-of-blogs-deleted.html), and Snapchat (http://en.wikipedia.org/wiki/Snapchat) which deletes contents after use.

Basis for Defining a National Web Sphere

The national part of the World Wide Web is here called the national web sphere. A web sphere is: “not simply a collection of web sites, but as a set of dynamically defined digital resources spanning multiple web sites deemed relevant or related to a central event, concept, or theme” [11], and here the theme is the country [2]. The national web sphere is in two parts: one part within the *ccTLD* and one part outside the *ccTLD* (*outside-ccTLD*) as illustrated in figure 1.

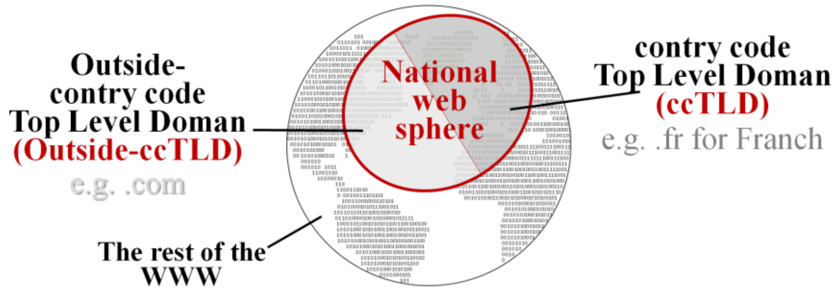

Figure 1: A national web sphere - and its parts

The *outside-ccTLD* is becoming increasingly interesting when defining a national web sphere. In the early age of internet collection, the focus for national institutions has been the within the country code Top Level Domain (ccTLD), but today there is a common awareness that there is a lot *outside-ccTLD* for a country. For example in France, it is estimated that materials from the .fr TLD only represent 33% of the total French web sphere ([13] page 9). An even stronger reason for the interest is that there is evidence that it is in continuous growth in size and change in contents of the *outside-ccTLD* [10]. One of the major reasons for this is that increasingly more information is published via what is often termed social media (e.g. Twitter and Facebook), other national TLDs (e.g. .se, .no) or other generic TLD (e.g. .com, .info, .nu and .org).

Defining a National Web Sphere has been subject for studies for different countries with different approaches. In 2008 the Czech web archive documented their suggestions in the paper “Identification and Archiving of the Czech Web Outside the National Domain” from 2008 [14].

When debating how to define a national web sphere, it has to be acknowledged that a country’s web sphere does not exist as such in itself. Therefore, it is important to reflect critically upon how a national web sphere definition is constructed. The claim here is that the following three perspectives must be represented to contribute to the final decisions on what is actually included in the web archive’s coverage of a national web sphere:

- *The Perspective of Collection* is usually represented by person(s) responsible for a web archive collection, and ensures that the web archive fulfils its purposes. These purposes are e.g. concerned with the requirements for collection scope and sustainable which can origin from public laws and regulations as for example a legal deposit law.

  The challenges in web archiving makes compromises necessary, and it is usually in this perspective that other perspectives as well as available resources (economy, labour or technology) are taken into account in order to make the final strategies.
- *The Perspective of Technology* is concerned with the computer science aspects giving technical possibilities and limitations as well as changes in these over time as a consequence of technological

evolution. The perspective covers technical aspect of collection, but also challenges with preservation and access, since these parts of a web archive are likely to be interdependent [8,15].

For collection of the national web sphere, an obvious area of interest lies where technological limitations are imposed on web archiving and harvesting. Examples that have caused challenges are Flash, streaming and java-script based webpages [9].

- *The Perspective of Scholarship* is concerned with the concept of the national web, and use of the national web for present and future research and audiences. Researchers, with this perspective, can contribute with input to decisions on selection of strategies by arguing for adjustments of collection strategies based on scientific needs to be able to answer typical research questions.

  In order for scholars to use the web archive, there are also many requirements for access to different facets of web archive material, e.g. viewing a webpage, linguistic analysis on the full web archive, limited material from a narrow time interval, or subject oriented parts.

The choice of these three perspectives follows a long tradition by which national libraries define what is national by involving both the *perspectives of collection* and *scholarship*. As regards the digital national library the *perspective of technology* needs to be included as well.

The Figure 2 illustrates that the *perspectives of collection* and *technology* are limiting the scope of the national web sphere, which relates to the resource limits (for collection) and technological challenges. On the other hand, the *perspective of scholarship* is seen extending beyond both of these in defining the national web sphere. The reasons for this are that researchers would like to have as much material to choose from as possible, and their interests may go beyond the national. However, this material may not be considered ’national' in the perspective of collection. None of the perspectives are illustrated to cover all, since we cannot claim to know all web material.

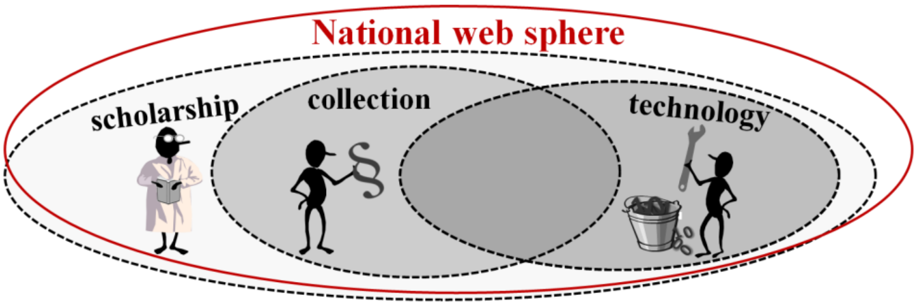

Figure 2: Union of the three perspectives

Compared to the *perspectives of collection* and *technology*, the *perspective of scholarship* is different in terms of time in the sense that the delimitation of a nation’s web is always made post festum in relation to the web archive. This means that in many cases the criteria for delimiting the nation’s web sphere will follow the criteria already used by the web archive. Therefore even after the interpretation of e.g. a legal deposit law, there can be discussion as to what to discard, and what is in danger of being missed.

As the very basis for a web archives’ existence is the use of a web archive, wishes for access are relevant for *the perspective of collection*. Although access is not directly linked to a web sphere definition, it has high influence on costs and technology platform, thus it may very well have significant impact when prioritizing resources and deciding on metadata and technology in *the perspectives of technology* and *collection*.

National Recognition Criteria

This section discusses general recognition criteria which can be used as basis for evaluating which and how to include criteria when defining what is within a national web sphere at given point in time, evaluated from the three perspectives.

The criteria originates from different sources: *known practices* that are already in use in different countries either in manual or automatic form, *former Czech suggestions* as mentioned earlier [14], *language recognition* practices from computer science used by e.g. for the Olympic game collection 2012 [1], *FRBROO reference parts* [5] and *brainstorm* on possible future practices listed from a brainstorm between the representatives[4] of the different perspectives. Each of the criteria was formulated generally in order to be able to fit any country. They roughly cover the following topics:

- ***ccTLD*** which is the traditional criteria based on a country's TLD
- ***Language criteria*** such as language recognition (e.g. N-gram analysis[3], special characters etc.)
- ***National icons*** e.g. corporate body, person, places, concepts or physical things like a famous painting
- ***Pointers to country*** e.g. use of standard country codes (in mail, phone numbers, or URL), IP-address in country, links to *ccTLD* etc.
- ***Events*** which for example could be a national election

There will be different possibilities for implementation and automation of these general criteria depending on the country (e.g. internet usage[5] and language) and present technology. This has also been the case for the WebDanica project looking at Denmark. Note that the criteria calculations are independent of the manifestations of data they are based on, e.g. in elements of a webpage. For example, text can be found in the URL, the html, in text from e.g. a PDF file, from an image, transcribed sound, in a video etc. Thus the criteria are independent of new sorts of elements that may appear on the internet in the future.

The combination of criteria is important in order to get an optimal automatic detection of material from the *outside-ccTLD*. Many of the criteria have shown out to be vague on their own in practice. However in combination with other criteria a vague criterion can strengthen automatic decision on whether a webpage belongs to a national web sphere. An example of a dedicated combination is the language recognition which becomes stronger if combined by recognition of distinguishable national words [12].

Analysis of the criteria showed out to help in optimization of costs in the WebDanica project. Originally, the project was supposed to run on 365 TB of data from Internet Archive. However, only the URLs, html and links were used in the calculation. Thus by deriving this information the amount of data could be reduced to 37 TB and consequently cost reductions.

In evaluation, representations of the three perspectives are crucial, since changes are very likely to be represented in the three perspectives. For instance, enhanced technology or limits for new technology (e.g. for sound, image, or emulation) will be represented by the *perspective of technology*, changes in funding and jurisdiction (e.g. public access, cut in preservation budgets) will be represented by the *perspective of collection* and trends and limits for present and future research (e.g. changed behaviour) will be represented by the *perspective of scholarship*.

[4] Group of the authors extended with other specialists.

[5] See e.g. http://www.internetworldstats.com/stats.htm

Discussion and Further Work

Some may claim that a national web sphere is not meaningful when considering the current level of globalization. However, without the idea of a distinct national web sphere, national cultural heritage will be lost.

Having general criteria, as defined in this paper, can help in defining how to collect (and construct) a national web sphere. However, having general criteria can also lead to an assumption that different national web spheres are readily comparable. In making such comparisons, it should be noted that there are many different local circumstances in any web sphere, for example the legal framework restricting collection strategies etc.

Another difference between national web archives might be the *perspectives of collection*, *technology* and *scholarship*. Particularly the perspective of collection may differ a lot due to legal or cultural differences, budgets and the formerly mentioned possibilities for collection of the national web sphere. The *perspective of scholarship* might also differ, for example due to different research cultures, or differences in the national culture regarding behaviour on the internet. An example of behaviour that differs between countries is the use of Facebook, which can be higher for some countries than others at different points in time. Another difference can be additional expenses, for instance if sound and images are included in search material.

Evaluation of selection strategies is needed, but there is at this stage no guideline for establishing such procedures. Similar issues arise when the *perspective of scholarship* is included in the decision setup, and whether it can rely solely on triggers for changes or should be considered periodically. Such guidelines still remain to be evolved.

Conclusion

This paper has presented an approach to reach a general definition of a national web sphere as the potential material for a national web archive, to cover national heritage, and be basis for future research. It has been argued that there are three perspectives that need to be taken into account in the decision making regarding strategies for automatic collection of the *outside-ccTLD* material, namely:

- The *perspective of collection* concerned with ensuring that the requirements for collection scope are fulfilled, covering legal framework, resources, needs and technology.
- The *perspective of technology* concerned with the technical possibilities and limits in connection with collection and sustainable access of the collection.
- The *perspective of scholarship* concerned with the concept and contents of the national web sphere for present and future research and audiences.

The paper has argued for the usefulness of a general list of criteria and mentioned a range of compromises to be taken from each of the three perspectives, in order to achieve the optimal possible automated collection practice for future web archive expansion.

Finally the paper has pointed to the importance of regularly reviewing of pre-conditions from the three perspectives, e.g. to catch changes in legal framework, in use of the internet and new technology to collect, preserve and give access to a web archive.

References

[1] Binns, A. 2013. *A look at languages in the 2012 Olympics web archive*. Available at http://aaron.blog.archive.org/2013/05/09/a-look-at-languages-in-the-2012-olympics-web-archive/ retrieved March 2014.

[2] Brügger, N. 2014. *Probing a nation's web sphere: A new approach to web history and a new kind of historical source*. Accepted for the 64th annual conference of the International Communication Association (ICA), Seattle, 2014.

[3] Chen, H.-H., Lee, Y.-S. 1994. *Approximate N-Gram Markov Model for Natural Language Generation*. Available at http://arxiv.org/pdf/cmp-lg/9408012.pdf, retrieved March 2014.

[4] Guy, M. 2009. *What's the average lifespan of a Web page?* http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page/, retrieved March 2014.

[5] International Working Group on FRBR and CIDOC CRM Harmonisation 2015. *FRBR object-oriented definition and mapping from FRBRER, FRAD and FRSAD (version 2.1). Eds. Bekiari, C., Doerr, M., Le Bœuf, P., Riva, P.* Available at http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V2.1_2015February.docx

[6] Internet Preservation Consortium (IIPC), Information available at http://netpreserve.org/ retrieved March 2014.

[7] Internet Archive, wide-00005 world wide web crawl. Information available at https://home.archive.org/~vinay/wide/wide-00005.html retrieved March 2014.

[8] Jurik, B. A., Zierau, E. 2011. *Different Mass Processing Services in a Bit Repository*. In: Proceedings of the Fourth Workshop on Very Large Digital Libraries (VLDL 2011), Berlin, Germany, pp. 11-18.

[9] Lazorchak, B. 2012. *Designing Preservable Websites*, Redux. Available at http://blogs.loc.gov/digitalpreservation/2012/02/designing-preservable-websites-redux/, retrieved March 2014.

[10] [Mjøs, O. J. 2012. M*usic, social media and global mobility: MySpace, Facebook, YouTube*. Routledge Advances in Internationalizing Media Studies.

[11] Schneider, S.M., Foot, K.A. 2006. *Web Campaigning*. Cambridge, MA: MIT Press.

[12] Shuyo, N. 2013. *Why is Norwegian and Danish identification difficult?*. Available at http://shuyo.wordpress.com/2012/03/07/why-is-norwegian-and-danish-identification-difficult/, retrieved March 2014.

[13] Stirling, P, Illien, G., Sanz, P., Sepetjan, S. 2011. *The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future*. In: proceedings of 77th IFLA General Conference and Assembly, San Juan, Puerto Rico, 2011. Available at http://conference.ifla.org/past-wlic/2011/193-stirling-en.pdf, retrieved March 2014.

[14] Vlcek, I. 2008. *Identification and Archiving of the Czech Web Outside the National Domain*. IWAW 2008, Aarhus, Denmark. Available at http://iwaw.europarchive.org/08/IWAW2008-Vlcek.pdf, retrieved March 2014.

[15] Zierau, E. 2012. *A holistic approach to bit preservation.* Emerald Group Publishing Limited, Library Hi Tech, Vol. 30 Issue: 3, pp.472 – 489