

Thinking Nationally with the Web

A Medium-specific Approach to the National Turn in Web Archiving

Name: Esther Weltevrede

Address: Jollemanhof 190, 1019 GW Amsterdam

GSM: 06 41306756

email: weltevrede@digitalmethods.net

Student number: 0357685

Supervisor: prof. dr. Richard Rogers

Course: Master thesis

Department: Media Studies

University: University of Amsterdam

Date: April 10, 2009

Abstract

Web archives face the important task of saving digital cultural heritage. Most Web archiving initiatives base their collection in the archives on a predefined list of URLs. The challenge, however, is not only to select relevant Websites, but also the prominence of Websites in their digital environment, including relations between Websites, their functionality and place in a larger entity.

There are two larger points made in this study. The first is a new way to think of Web space. A medium-specific approach to Web spaces as ordered by 'technical arrangements' is introduced. There are a number of technical arrangements on the Web, including Web archives and search engines that order the once universal cyberspace in distinctive Web territories. In this piece I call attention to arrangements that order Web content and users along national or linguistic lines, and more specifically the 'national Webs.'

The second focuses on one specific type of technical arrangement: the Web archives. It strives to find out how and why current Web archives look as they do. Two Web archiving projects are looked at in more detail. The first project, which strives to save the entire Web since 1996, the Internet Archive, is compared to the Web archive of the Royal Library of the Netherlands (the KB) that started archiving a selection of Dutch Websites in 2006. When Web archivists think of 'time' they usually refer to the creation date of documents, here, the time from which the Web archives emerge is discussed. The hypothesis is that Web archives are shaped by the period and spirit of their creation, mirroring dominant thoughts as well as technical developments. However, it was found that the dominance of the institutional context from which they emerge should not be underestimated.

Building on the two larger points of this study, the effort is to contribute to archival theory and practice with collection techniques. The proposed techniques are ways to start thinking about saving the dynamic context of Web documents.

Keywords

national Webs, media of location, cyberspace, Web archives, digital methods, technical arrangements, technical indicators

Table of Contents

Abstract	3
Keywords	3
Table of contents	5
Acknowledgements	6
Introduction	7
Part 1: A Technical Approach to the National Turn	13
1. The Digital Methods Initiative	15
2. Thinking National with the Web?	17
3. The Webs as Media of Location	24
Cybergeography	29
The Yahoo! Case: IP-to-geo	30
The Domain Name System	31
Part 2: Archival Principles and the National Turn	41
4. Theories of the Archives	42
Archival Principles and Practices	44
The Web as an Ephemeral Archive	47
5. Archiving Cyberspace: the Internet Archive	49
Cyber Spatial Concerns	50
Cyberspace in a Box	52
The Web Archive and the Digital Library	54
Heritrix, Metadata and the Wayback Machine	54
The Utilitarian Dream	56
The Academic Archives	58
Legal Issues and Robots.txt	59
The Internet Archive in 2009	60
6. Archiving the National Web	62
Web Archiving Models	67
KB's Definition of the Dutch Web	71
The Selective Approach	73
7. The Order of Things in the Digital	76
Medium-specific Collection Techniques	82
Conclusion	84
References	87
Web References	92
Figures References	98
Appendices	102
Appendix A	102
Appendix B	108
Appendix C	110

Acknowledgements

My heartfelt thanks go out to my friends and colleagues at the Digital Methods Initiative and Govcom.org for their inspirational thoughts and comments on my research project. Part of this thesis was presented at the first DMI writing seminar in January 2009. I would like to thank the participants, Erik Borra, Andrea Fiore, Anne Helmond, Sabine Niederer, Richard Rogers, Michael Stevenson, Laura van der Vlies and Marijn de Vries Hoogerwerff for their comments and input.

Thoughts on the national Webs were developed over the last two years; in particular by specific projects such as Mapping Palestinian Cyberspace, 2007/2008, and The Govcom.org Jubilee project Palestinian Cyberlands 2.0, 2008. My thanks go to the participants of several DMI and Govcom.org projects, with additional contributions from Anat Ben-David, Mike Dahan, Isabelle Daneels, Marieke van Dijk, Ganaele Langlois, Astrid Mager, Koen Martens, Rosa Menkman, Andrei Mogoutov, Rafal Rohozinski, Charmaine Stanley and Auke Touwslager. I would like to thank Jose van Dijck for her contribution to my PhD proposal, in which this study will find its place.

Special thanks goes out to my supervisor Richard Rogers for his inspiring meta-view, Erik Borra for his structure, insights and loving support while writing this study, Marguerite Lely for her warmly appreciated language expertise and my PhD-friends Sabine Niederer and Michael Stevenson for peer-reviewing.

Introduction

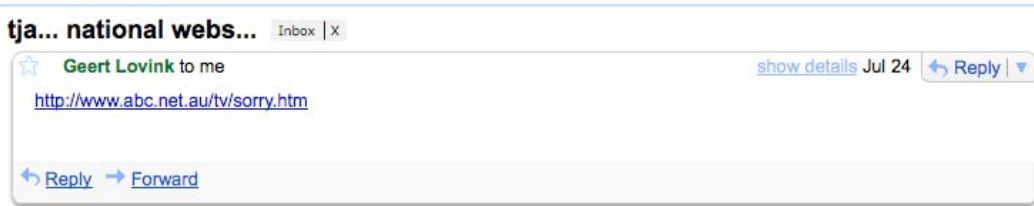
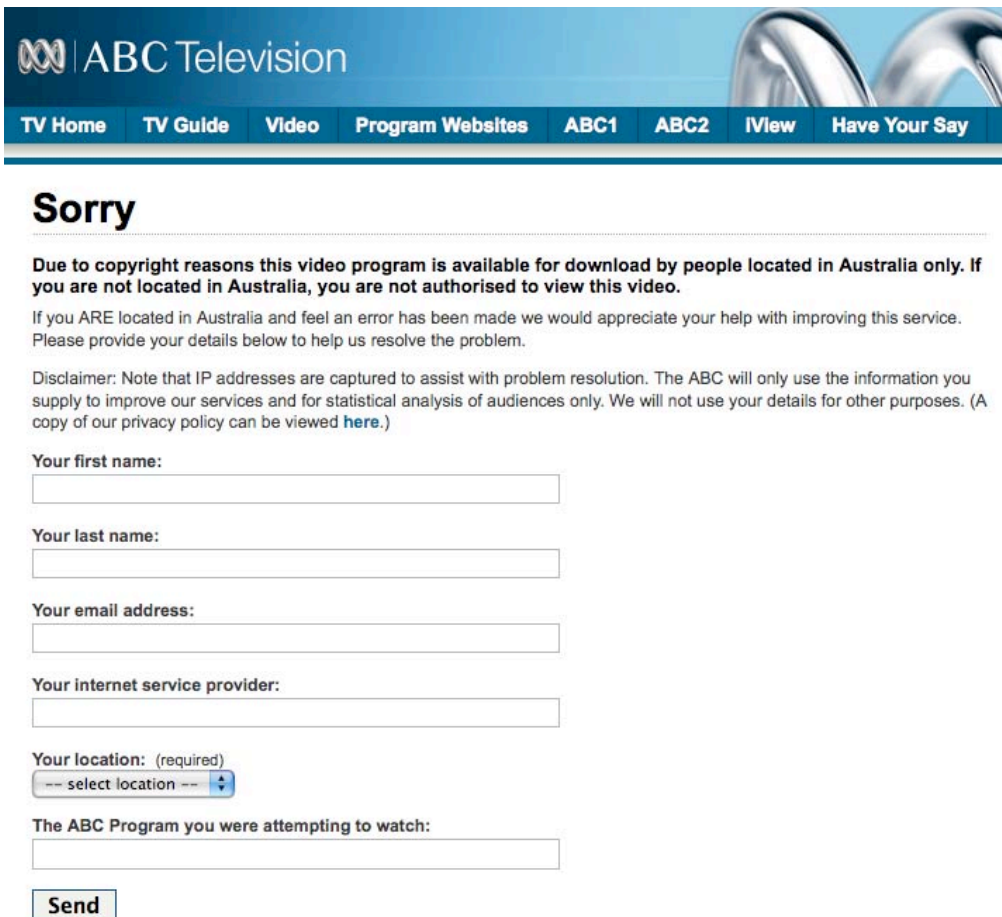


Figure 1. “tja... national Webs...” email, 2008

We have all become familiar with search engines that by default redirect us to a regional version. Increasingly, however, we are confronted with geographical issues at other places on the Web as well. The reasons for this vary and technically speaking they are implemented in different ways. One of my professors, Geert Lovink, sent me an email about the ABC.net.au Website, which he could not access because of his geographic location (figure 1). ABC Television states that because of national intellectual property legislation content is blocked for those outside Australia (figure 2). ABC enforces this legislation through computers' Internet Protocol (IP)-addresses, assigned to them by Internet service providers (ISP). This numerical identifier of a computer is cross-referenced with IP-ranges assigned to geographical regions or companies; they are used to determine where users are geographically based.¹ This system, however, is not fully flawless, so ABC has an accompanying form to be filled out if you are based in Australia (figure 2).

¹ To generate a list of IP-ranges for any country, see Find IP-address, 2009



ABC Television

TV Home TV Guide Video Program Websites ABC1 ABC2 IView Have Your Say

Sorry

Due to copyright reasons this video program is available for download by people located in Australia only. If you are not located in Australia, you are not authorised to view this video.

If you ARE located in Australia and feel an error has been made we would appreciate your help with improving this service. Please provide your details below to help us resolve the problem.

Disclaimer: Note that IP addresses are captured to assist with problem resolution. The ABC will only use the information you supply to improve our services and for statistical analysis of audiences only. We will not use your details for other purposes. (A copy of our privacy policy can be viewed [here](#).)

Your first name:

Your last name:

Your email address:

Your internet service provider:

Your location: (required)

The ABC Program you were attempting to watch:

Figure 2. Sorry, ABC Television 2008

YouTube.com's "This video is not available in your country" has a similar yet slightly different national Web story for blocking content in specific areas of the world (figure 3).



This video is not available in your country.

Figure 3. This video is not available in your country, YouTube 2008

YouTube uses similar technology to block content in specific geographic regions. The discussion on YouTube's community help forum, however, shows that users do not understand why and how videos are blocked (figure 4). In the Netherlands, the video that is allegedly blocked in at least Kuwait can be viewed. The URL http://www.youtube.com/watch?v=S88rkpPu8_g, however, is by default redirected to http://nl.youtube.com/watch?v=S88rkpPu8_g. The URL's added prefix 'nl' already indicates that YouTube is able to serve content nationally or regionally. This blocking of videos has noth-

ing to do with governmental censorship like in countries such as Pakistan or China where governments regularly have access to content blocked by ISPs and results obfuscated by search engines.² A blog post frequently referenced to explain why YouTube videos are unavailable is *Digital Inspiration's* "YouTube Video Not Available in Your Country? How to Watch Blocked Videos" (Agarwal 2008):

If your computer's IP-address falls outside that geographic region, YouTube will display an error saying 'This video is not available in your country' - this message has nothing to do with censorship, it's the owner of the video clip who could be limiting access.

Whereas ABC claims that content is blocked because of national copyright law, YouTube restrictions might be due to national intellectual property legislation, but it can also be any or no reason since any individual uploader can determine where a video is available. Media companies, such as broadcasters and record labels, use marketing reasons, or are bound by complicated licensing or legal issues. A video may be legal in some countries, but not in others (e.g. videos denying the Holocaust are illegal in Germany). The technical apparatus used for serving content nationally, the IP-addresses, is the same as with ABC. The reasons for blocking content, however, are different. This is confusing for video site users, as the innumerable discussions online about how and why YouTube blocks content show.

² For research on Internet censorship in various countries see OpenNet Initiative 2009 and Reporters Without Borders 2008.

Community Help Forums

Discussions > Bug Reports & Issues > Video not available in my country

Options

5 messages - [Collapse all](#)**SourcererKhadgar** [View profile](#)[More options](#) Oct 17, 12:36 am

I have viewed this video:
http://www.youtube.com/watch?v=S88rkpPu8_g
 many times before and now suddenly I'm getting a message that it's not available in my country.

Why? What's changed?

Thanks,
 Nikola

[Reply](#) [Forward](#)
MerryChristmas2You [View profile](#)[More options](#) Oct 17, 6:24 am

unknown, the message you got is usually based upon your IP address.

<http://www.labnol.org/internet/video/youtube-blocked-video-not-availa...>

On Oct 16, 5:36 pm, SourcererKhadgar wrote:

- Show quoted text -

[Reply](#) [Forward](#)
SourcererKhadgar [View profile](#)[More options](#) Oct 17, 6:40 am

Hello,

By following the instructions on the site you gave me, I got the message that the video is no longer available. Were YOU able to watch the video?

Nikola

On Oct 17, 6:24 am, MerryChristmas2You wrote:

- Show quoted text -

[Reply](#) [Forward](#)
MerryChristmas2You [View profile](#)[More options](#) Oct 17, 7:32 am

i can watch the video just fine without getting any message.
 i live in the U.S. , midwest.
 maybe that makes a difference or it's just a bug of some kind.

On Oct 16, 11:40 pm, SourcererKhadgar wrote:

- Show quoted text -

[Reply](#) [Forward](#)
bashar80 [View profile](#)[More options](#) Oct 18, 11:26 am

Hi,

I am in Kuwait and I also can't watch your video. Where are you living, and can you watch this one?

<http://www.youtube.com/watch?v=rXEEKP3g1gY>

The trick by MerryChristmas2You worked, but it's not sufficient to me. I need the video to be embedded and searchable.

In my case, this specific video up there was first banned for audio track copyright, I replaced it with AudioSwap, now it's not working in all countries. Using South Korea proxy though, I was able to play it!

On Oct 17, 7:40 am, SourcererKhadgar wrote:

- Show quoted text -

[Reply](#) [Forward](#)

End of messages

Home

Discussions

[About this group](#)[Join this group](#)

Figure 4. 'Video not available in my country,' YouTube Discussions 2008

In the above-mentioned stories from the Web three technical arrangements, defined as systems that order Web content by technically defined measures, are in place. They are: a search engine, a broadcasting site, and a video platform. They all order and subsequently serve Web content with location as an organizing element, demonstrating that the Web can be seen as media of location. All three use the same technology to define the borders of national Webs. They gauge the nationality of the users (or their computers) in the same way. Reasons for ascribing a nationality to content, however, are different: ABC defines its content as Australian to adhere to national intellectual property legislation while YouTube delegates the definition of the nationality of content to the uploader. The differences between defining the nationality of users and content to some extent reveals how each of the arrangements and related devices on the Web work. It is important to realize that there are many ways to think national with the Web; each of them is implemented via a specific technical apparatus, such as the domain system. A Web search engine uses IP-addresses in a different way than a video platform or a single Website.

It seems unusual to think of the Webs as a media of location since the notion of cyberspace is so global. This study has a double aim. In the first place, a medium-specific approach to look at Web spaces as 'national Webs' is introduced by placing it in the context of other conceptualizations of Internet spaces. This approach privileges thinking of national Webs in terms of 'technical arrangements' that configure cyberspace along national lines. It tries in particular to react in distinctive ways to calls of thinking about the medium that are specific, by claiming that the Webs are media of location.

Second, one particular technical arrangement underscoring the national turn is analyzed: Web archives. Web archives face the important task of saving cultural heritage for posterity. Building on Foucault and Derrida, the shape of the archive constrains and enables what can be known with the archives. Moreover, the technical methods that are used in the archiving process register as well as produce the object of collection. The first project that strives to save the entire Web, the Internet Archive, is compared to the Web archive of the Netherlands that started archiving a selection of Dutch Websites. The effort here is to strive to find out why the archives look as they do. The hypothesis is that Web archives are shaped by the period and spirit from which they emerge. The archivists' approach to the object of collection shape the archive. Since the first initiative that began archiving in 1996, the Internet Archive, a number of projects have emerged that archive with a national focus. The Web is however not simply organized along national lines, but rather technological solutions need to configure the Web as such.

After all, the word "archive" is derived from the Greek ἀρχή (arkhē), meaning government or order (compare an-archy or mon-archy). By considering Web archives as arrangements constructing national Webs, we can gain insight in how they have to handle technical apparatuses and technical arrangements enabling and constraining them to deal nationally with the Web. Thinking about the Webs as media of location, and more specifically as arranged along national or linguistic lines, is examined by looking at texts and images produced by Web archiving institutions selecting and ordering Web content as such. In this study I explore the origins of the Web archivists' technical choices. Considering all possible ways to technological define a national Web for a given country,

what technological basis do national Web archivists use? And, given their methods, what technical arrangements do archivists create?

Lastly, the novel approach to Web space is used to make a contribution to the field of Web archiving. With reference to traditional archival principles, medium-specific techniques for the collection process are proposed to preserve parts of the Web that will otherwise be lost.

Part 1: A Technical Approach to the National Turn

1. The Digital Methods Initiative

This study is placed in the context of the Digital Methods Initiative (DMI), and strives to contribute to it. The Digital Methods Initiative aims to develop a theory to recognize the existence and importance of the 'natively digital,' a category of digital objects or 'building blocks' of the Web, including the link, top-level domain (TLD) and IP-address. The initial question is: what methods are appropriate when the object of study has changed so dramatically? The underlying assumption of this question is a novel way of thinking about the Web as well as the development of skills to understand how the medium works. These technical insights are translated into methods and tools researching the organization of the Web from within (Rogers 2008a). One of its core assumptions is that there are Webs within the Web, plural and somewhat different. In the eyes of DMI, Google's Web is not Yahoo!'s.

The Digital Methods Initiative sprung out of the New Media program, forming part of the Media Studies department at the faculties of humanities at the University of Amsterdam, and Govcom.org, an Amsterdam-based foundation dedicated to creating and hosting political tools on the Web. Supervised by prof. Richard Rogers New Media students, Web researchers, designers and programmers cooperate to develop new methods for novel study objects. Hereafter the DMI is placed in the context of other areas in touch with new media, the digital and computing in the faculty of humanities.

The DMI has similarities with software studies, which is a relatively new umbrella concept for scholars choosing software as a new object of study. They consider software studies either as a new current within media studies or as a new, distinct field in which new media are the object of study, such as cyber culture, Internet studies and digital culture. The former approach does not attempt to start a new field of study, but instead calls for new theories of software in areas that "have not historically 'owned' software," such as media studies, but could lead to a new approach to software with critical perspectives on politics, society and matter (Fuller in Helmond, 2008). The latter considers software studies as a new intellectual paradigm, distinct from areas such as media studies (Manovich 2008).

Software theorist and artist Matthew Fuller considerably advanced software studies with his books *Behind the Blip: Essays on the Culture of Software* (2003), *Software Studies: A Lexicon* (2008) and the *Software Studies Workshop* at the Piet Zwart Institute in Rotterdam (2006). The Lexicon presents a broad and fascinating overview of next generation programmer-theorists contributing to software studies from their own perspective, offering an input from fields, which traditionally have no direct link with software, like philosophy, history, or visual culture studies.³ Taking the technical composition of digital systems as point of departure, the collection tries to move beyond considering software as a tool, or 'something that you do something with.' Software is not 'neutral': Fuller calls this the 'ideological' layer, which relies upon an understanding of the materiality of software being operative at many scales (2008: 6). Software studies differ from many publications in the area of new media, which mainly focus on content when describing phenomena like the Internet or

3 With contributions from Jussi Parikka, Wendy Chun, Florian Cramer, Warren Sack, Adrian McKenzie, Nick Monfort, Friedrich Kittler, Olga Goriunova, Alexei Shulgin and Graham Harwood,

games. According to Fuller, software studies emphasize the neglected aspect of computation, which involve virtuality, simulation, abstraction, feedback and autonomous processes (2008: 6). The lexicon tries to describe software studies, what they trigger and what they can be linked to, and in doing so the lexicon offers multiple entry points into the field. Instead of monitoring users behind their screens, it aims to map the conjunction where “computation meets with its ostensible outside (users, culture, aesthetics) but is not epistemically subordinated by it” (2008: 6)

New media and software theorist Lev Manovich coined the term ‘software studies’ in *The Language of New Media* (2001), where he called for a new approach to the object of study:

New media calls for a new stage in media theory whose beginnings can be traced back to the revolutionary works of Robert Innis and Marshall McLuhan of the 1950s. To understand the logic of new media we need to turn to computer science. It is there that we may expect to find the new terms, categories and operations that characterize media that became programmable. From media studies, we move to something which can be called software studies; from media theory — to software theory.

In *Software Takes Command* (2008) he nuances the importance of computer sciences as primary focus in software studies and redefines the challenges to “investigate both the role of software in forming contemporary culture, and cultural, social, and economic forces that are shaping development of software itself” (2008: 5). In other words, software is considered to be a layer that permeates all aspects of contemporary society. Among others, the book builds on *The New Media Reader* edited by computer scientist Noah Wardrip-Fruin and digital media scholar Nick Montfort (2003), which defined the intellectual framework for the historical study of software. The New Media Reader did not explicitly use the term ‘software studies,’ but proposed a new model for thinking about software by bringing together important texts by scientists and artists, including Jorge Borges, Vannevar Bush, Ivan Sutherland, Ted Nelson, and Douglas Engelbart. Manovich started a particular path through the conceptual history of media computing from the early 1960s until today, by drawing the genealogy of cultural software. He focuses on ‘content creation’ software and the systematic investigations of its roles in cultural production. In other words, to what extent have interfaces and the tools of content development software reshaped and are still shaping the aesthetics and visual languages applied in contemporary design and media (2008: 21)?

It is crucial to notice that these currents within software studies so far focus on the new object of study - software - but do not have an approach of their own: methods are imported from other fields, including computer science, philosophy, history and visual culture. With a focus on the relation between software and the Web, the DMI contributes to the study of this new object by developing methods that are medium-specific, by combining an empirical approach to digital media and culture with a new media theoretical approach. ‘Medium specific’ is defined here as studying the medium’s native structures, objects and dynamics, which have not existed before and outside the digital, with methods that mimic and thrive on its native ontology.⁴

⁴ The ‘medium specific’ approach is different from for example ‘remediation’ as approach to the object of study. Remediation allows thinking of the same medium in a radical different way, i.e. continuity in the evolution of media where each medium takes over characteristics of a previous medium (Bolter and Grusin, 1999).

It may be argued that imported methods do not fit the medium. New media environments - and the software-makers - implemented the medium algorithmically, in ways that do not agree with the familiar schools of thought and methods. DMI tries "not simply to import well-known methods - be they from humanities, social science or computing. The focus is rather on how methods may change, however slightly or wholesale, owing to the technical specificities of new media" (Rogers 2008a). The natively digital is recognized as an object of study, and methods for studying the natively digital include crawlers and scrapers so as to study the organization of the Web from within. Web epistemologist Richard Rogers introduced this particular approach to the medium in his *Information Politics on the Web* (2004); he shows that the Web has its own mechanisms to determine the value and relevance of information.

2. Thinking National with the Web?

“We have many countries and many laws and just one Internet”

Heather Killen, Yahoo! vice president, 2000

It might seem unusual to think of the Webs as a media of location due to the dominance of the cyberspace notion. Prior to 2000 thinking was dominated by views of the Internet as one single space, separate from reality. Cyberspace, coined by William Gibson in *Neuromancer* (1984), refers to a virtual reality mediated by communication networks. This cyberspace privileges thinking of the Internet as a visual representation of data. Cyberspace informs ideas of an Internet that are technically indifferent to the geographical location of its users and their content, paralleling ideas of disembodiment, equality and identity play (Chun 2006). It moves the Internet beyond information on a screen, thus making it an inhabitable and navigational place. With *THE MATRIX* (1999), the Wachowski Brothers brought this idea to its ultimate imagination, reducing everyone and everything to discrete zeros and ones, visible once jacked into the system (figure 5). Digital rights activist John Perry Barlow used the term cyberspace to refer to the social spaces of the Internet (1996). Visual or social, the crucial sense of cyberspace is that it is a space disconnected and distinct from reality. The predicted final point of cyber spatial thinking is a disassociation of social life online from physical reality. Hereafter the approach to national Web spaces created by ‘technical arrangements’ is introduced. It is placed in context by discussing authors with different views on Web spaces and the end of cyberspace.

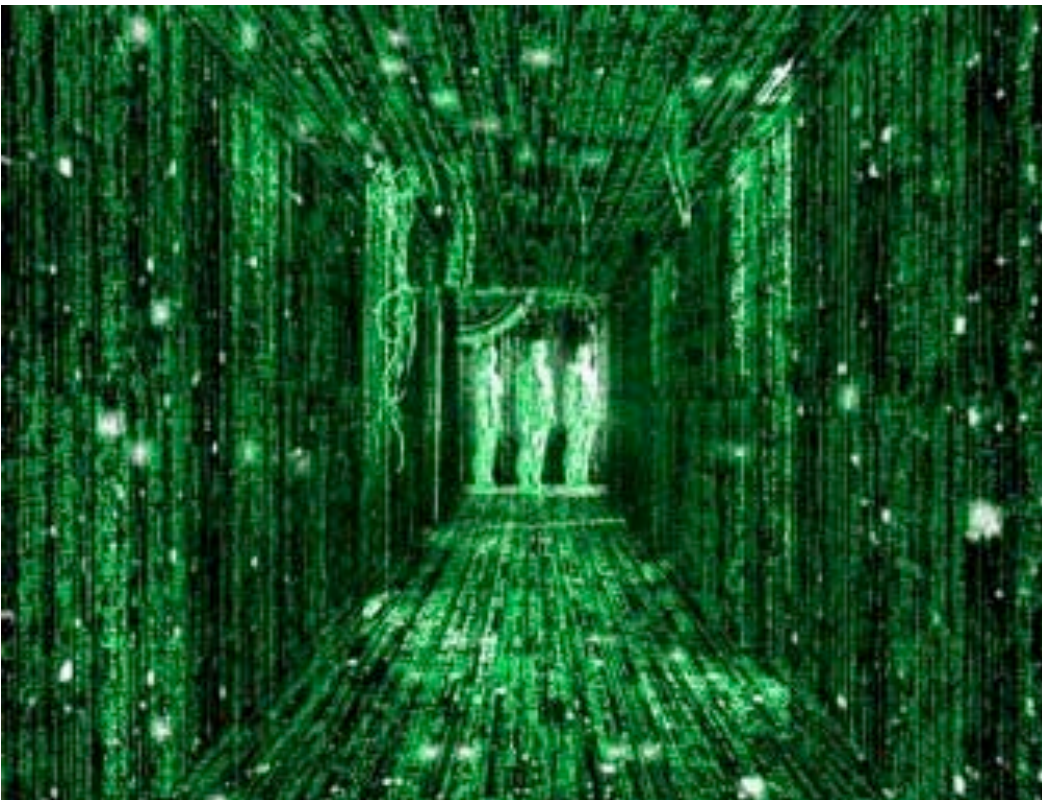


Figure 5. Cyberspace I: *The Matrix*

In *Control and Freedom* (2006) critical media theorist Wendy Chun argues that there is no 'space' in cyberspace when she positions the idea of cyberspace in popular media. With a study of William Gibson's *Neuromancer* (1984) and Mamoru Oshii's *GHOST IN THE SHELL* (1995) at hand, she examines two different versions of cyberspace: Gibson's cyberspace as something we jack into, and Oshii's cyberspace as something that jacks into us. Both notions of cyberspace make it possible to think of cyberspace as an information space that is identifiable yet unable to locate. Traditional assumptions about maps and space are put upside down: "cyberspace as world of disembodiment reduces locations and people to information, while at the same time it creates new information-based geographies" (2006: 115). With this approach to cyberspace as a non-space, the Communications Decency Act (CDA), section "Findings of Fact," moved cyberspace out of the realm of popular culture and made it into a legitimate communications medium (figure 7). The term cyberspace was chosen over the word Internet, because in this way configurations such as local area networks and bulletin board systems, which do not necessarily link to the Internet, could be included (Chun 2006: 43). This Act introduced 'space' to cyber spatial thinking by delineating different 'areas' within cyberspace as a communication medium (e.g. e-mail, World Wide Web, Internet Relay Chat).

Thinking about the Internet in terms of cyberspace has known a number of symbolic turning points which can be summarized as the 'national turn'. One field of study describing the national Web is the emerging field of virtual ethnography. The national Web is not addressed so often, but if so it follows the lines of: "we need to treat Internet media as continuous with and embedded in other social spaces" (Miller and Slater 2000: 5). The virtual methods approach to the Web directly reacts against the universal idea of cyberspace. The e-social science researchers from the U.K. Virtual Society research program of the late 1990s (Woolgar, 2002), questioned the then dominant view of the Web as a placeless cyberspace where everyone is equal and differences in race, class, and gender are overcome. Virtual methods may be seen as an exercise to measure the new technologies' impact on society and, more specifically, on the user. By 'visiting the ground'⁵ the Web is made comprehensible as an important social and political space. Put differently, the Web's embeddedness in society is sounded out on a variety of users in all sorts of cultural and social contexts offline. On of their most well known notion, the 'digital divide,' entails that access to cyberspace is not equally distributed across the globe (figure 6).⁶ This contribution is important as it shows that cyberspace's empowering promise is not evenly distributed in all geographical regions.

⁵ In this context 'ground' refers to the offline reality. With a medium-specific approach, the phrases 'grounding' and 'digital grounding' I explain in chapter 3. The Webs as Media of Location refer to locating content or users on the Web. The former refers to geographically locating, the latter to the relative location or position of content or users in a certain Web space.

⁶ Internet usage data in this graphic is gathered from Internet World Stats 2009.

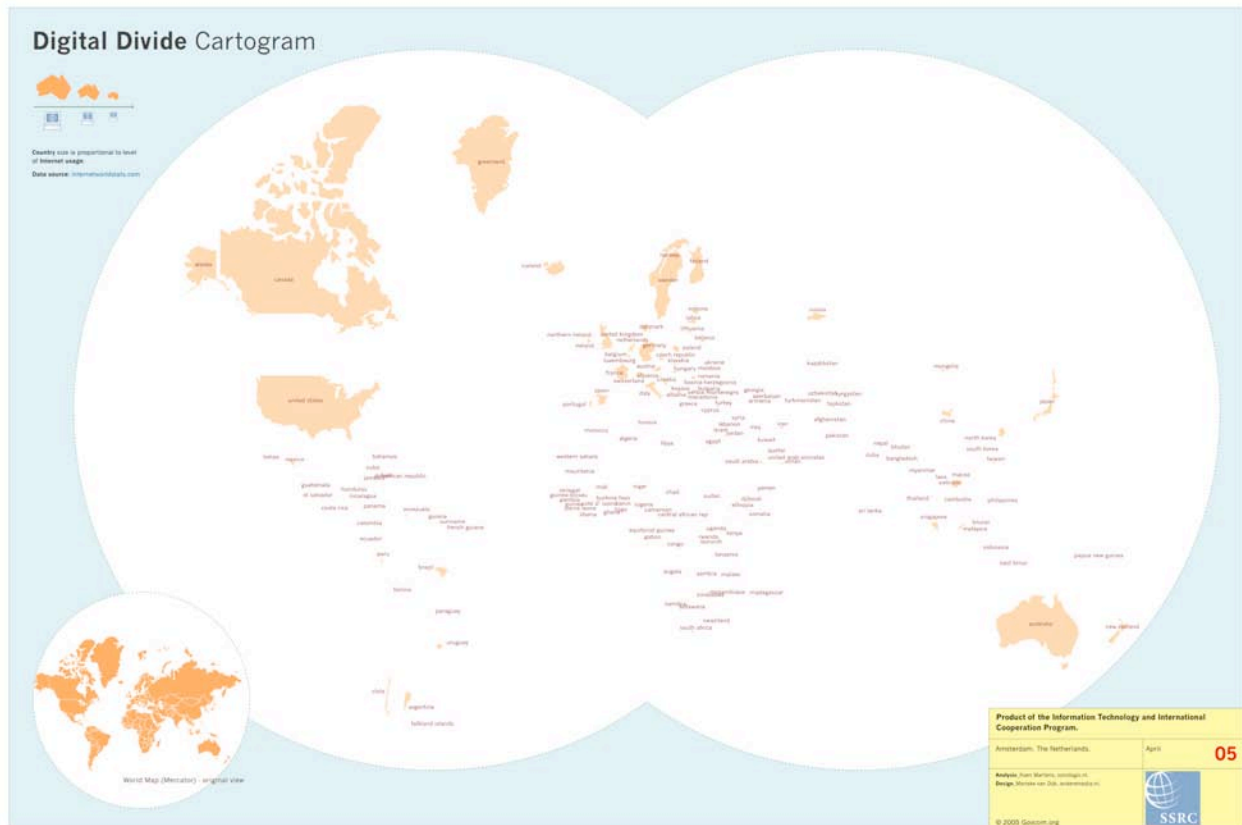


Figure 6. *Digital Divide Cartogram*, Govcom.org, 2005

Digital divide scholars as well as the Virtual Society? program through their methods look at how they the online creates accounts from the offline (Hine, 2005). The national Web, for its part, is considered to be understood by visiting the ground. Here, the aim is to turn this around: instead of studying the Web from a separate real space offline, the Webs are considered as national reconstructions from within the medium itself. This approach can thus be viewed as a reaction to the offline approach by studying a variety of locative technical objects and attempting to digitally ground the social element in the Web itself.

In *Here Comes Everybody* (2008), Clay Shirky's approach to the end of cyberspace coincides with the increasing public spread of the Internet. In the early stages of the Internet the average user interacted with different people online and offline. The idea of cyberspace made sense when the Internet population consisted of a few million users, and social relations online were really separate from those offline, because the people you met online were different from the people you met offline, and these worlds would rarely overlap. The separation between online and the real world, which is key to cyber spatial thinking, was an "accident of partial adoption" (Shirky 2008: 195). Studying national Webs with this approach entails a focus on the number of users per country and studying the correlations between social relations online and offline. It approaches space in terms of the overlap between social spaces online and offline.

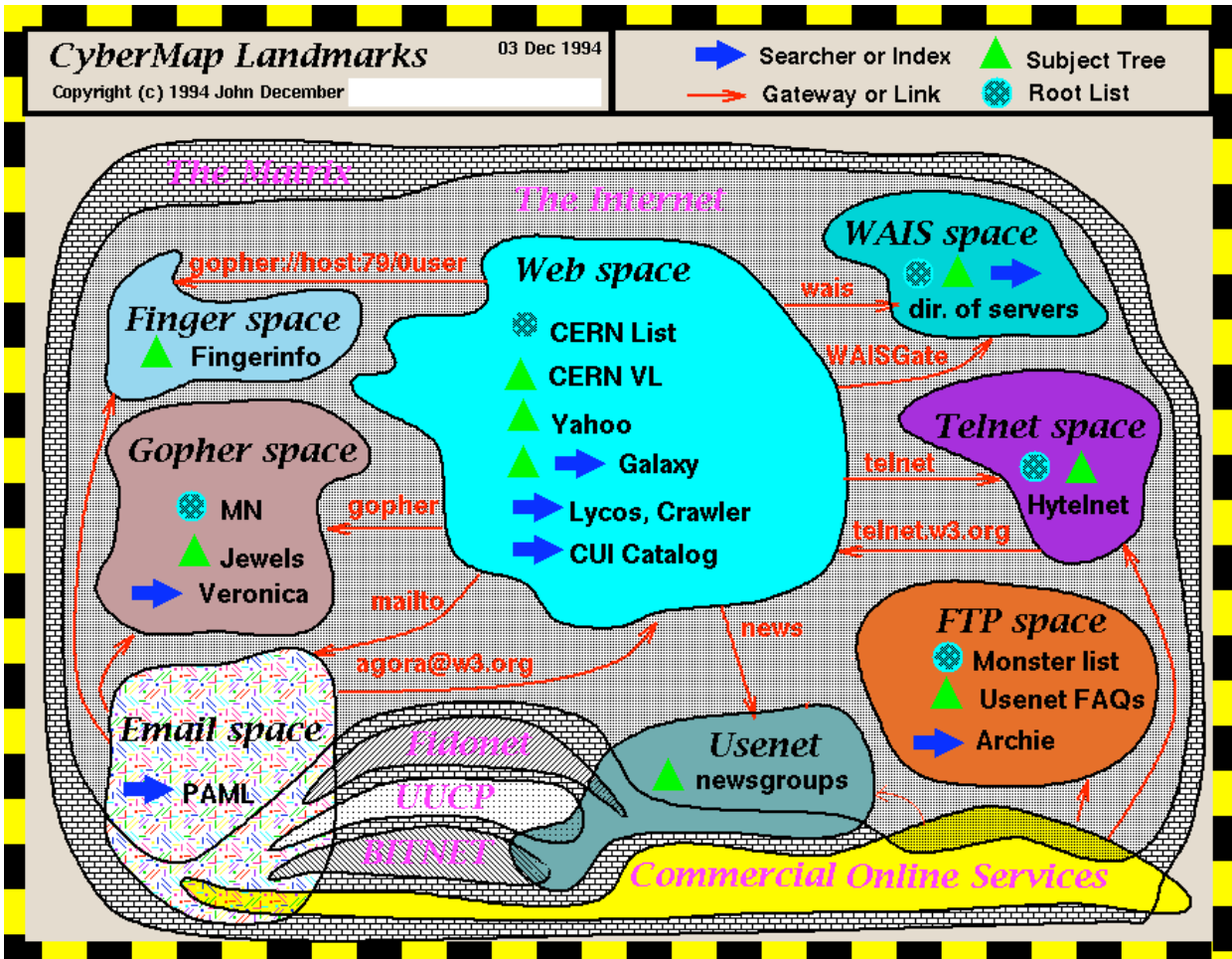


Figure 7. Cyberspace II: CyberMap Landmarks, John December 1994

In his master thesis "Measuring National Borders on the World Wide Web" (1998) Internet researcher Alex Halavais describes a national Web approach to the end of cyberspace. He makes a distinction between the concepts of 'nation' and 'state.' The nation is defined as "a group among which communication flows are strongest and differentiated from other nations by a relative lack of communication." State, on the other hand, "corresponds to the institutions and systems of control and power" (1998: 52). Nation is a social construction circumscribed by communication flows, whereas state is a governing system defined by a territory. With Halavais' approach to national Webs, space can be studied by measurable communication 'flows.'

Internet critic Geert Lovink theorizes the demise of cyberspace with the rise of Web spaces separated by languages. This approach introduces means to study national Webs by the clustering of online social communication based on a shared language. "The good thing about the 'national webs' [...] is that they are confined spaces. This is finally a big step away from the utopian 1990's way of thinking about cyberspace as a profoundly global space and discourse on what is really global and whether global means English. In these new spaces language plays a strange role, because it also facilitates the democratization of the medium itself" (Lovink 2009). In "The Polyglot Internet," co-founder of Global Voices Online, Ethan Zuckerman, thinks along the same lines as Lovink. Although Zuckerman takes a cyberspace-nostalgic point of view: "As the Internet becomes

less of a global, shared space and more of a Chinese or Arabic or English space, we lose incentives to work together on common, compatible frameworks and protocols. We face the real possibility of the Internet becoming multiple Internets, divided first by languages, but later by values, norms and protocols" (Zuckerman 2009). From the point of view of critical Internet culture, Lovink disagrees with Zuckerman by stressing the importance of the democratization of the medium with the rise of language Webs. Protocols and codes are currently predominantly in English and in the current state of the Internet, "English as a technical language can now be surpassed by other languages" (2009). The democratization of the medium would entail that the global decision making process concerning the Internet architecture and the protocols becomes accessible to other languages.

There are a number of approaches that theorize the end of cyberspace. They all witness a similar trend in conflict with the dominant view of a global shared space. All approaches opt for a new way of thinking about Web space, one that privileges thinking about Webs in plural and as delineated spaces. Each of the approaches, however, has a different theory how we should approach these national Web spaces. Thinking in terms of access, users, flows or language ends up with different boundaries of national Webs in each case. The national Web of the Netherlands e.g. is therefore a multiplicity of Webs, depending on the approach. These various approaches studying Webs within the Web contribute to the so-called 'national turn'.

In this study a novel approach to Web spaces is introduced, and in particular national Web spaces. The above-mentioned approaches think of the end of cyberspace in terms of users, communication flows and language. Another way is to think about it in terms of its technical organization. The approach proposed is medium-specific, which means that it calls attention to the objects, structures and dynamics of the Web that did not exist before and outside the digital. Before discussing the national Web with a medium-specific approach, the last cyber spatial thinker is addressed, the one thinking about cyberspace in technical terms. In *Protocol: How Control Exists after Decentralization* Alexander Galloway theorizes the protocol layers of the Internet as the management style to control the Internet, more specifically the Transmission Control Protocol/Internet Protocol (TCP/IP) (figure 8).⁷ Lower-level layers are encapsulated in higher-level layers. In other words, the Internet is considered to consist of technologies built on top of other technologies. The TCP/IP approach, like cyberspace, considers the Internet to be one single medium and starts out from its technical infrastructural reality and can therefore be defined as a technical approach to cyberspace.

⁷ TCP/IP are protocols that govern the Internet on a technical infrastructure level and are documented in Request For Comments (RFCs). TCP/IP is part of four layers of the Internet suite of protocols

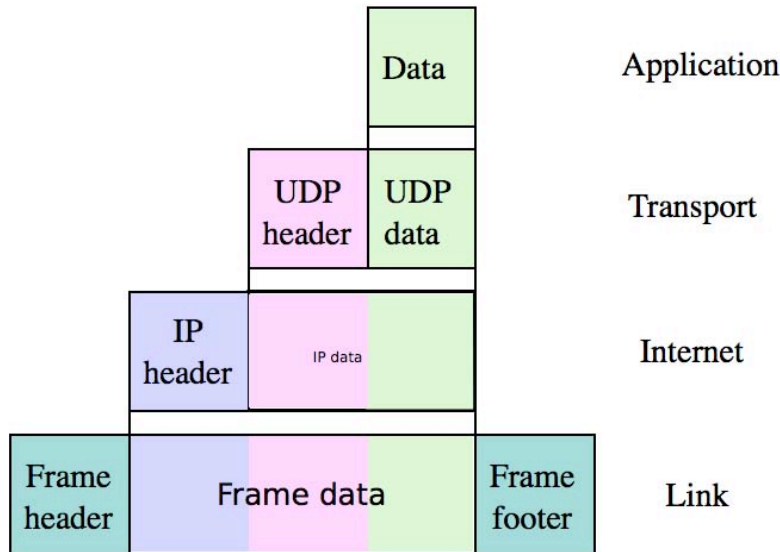


Figure 8. TCP/IP layered model of the Internet infrastructure

Placing the medium-specific approach to national Web research in the larger area of media analysis, an important concept in thinking about the 'national' is the 'imagined community,' developed by Benedict Anderson (Anderson 1991). The nation is a socially defined community, in the end imagined by those who consider themselves to be part of it. An imagined national community differs from an actual national community, as there is no direct communication between its members. It is a mental image, imagined by its members through shared media such as national cinema. With the medium-specific approach to national Webs, however, the national community is not as much imagined through shared media. Rather it is a shift in focus from the national as an imagined community to the national community as a technically mediated, place related construction from within the medium itself. The question then is not how the national element is imagined through the Web, but how we can find, map and diagnose it.

In her study on the Palestinian Web, Web researcher Anat Ben-David takes a medium-specific approach to study the national Web. She moves the discussion of the 'imagined' in cyberspace beyond imagined communities and identities, and goes to imagined places and geographies. She presents the Palestinian Web as an example of both 'imagined nationhood' and 'imagined statehood.' She argues that, "the Palestinian cyberstate bypasses the geographic reality on the ground and provides both continuously demarcated space and communication means for advancing public debate, polity, and establishment of the kind of statehood the anticipated Palestinian state wishes to realize on the ground" (Ben-David 2008). The .ps top-level domain grants autonomy to the Palestinian state on the Web. Moreover, it provides the opportunity to study social relationships within the delineated Webstate and between it and the ground. The .ps Webstate is an existing demarcated space that can be ruled via technology, while the state on the ground is imaginary.

Ben-David defines border of the Palestinian Web in terms of the top-level domain .ps. Using this technical apparatus to study Web space demonstrates that, although the Internet is a fairly distributed medium, there are technical systems that stratify and order the universal cyberspace.

There are a wide variety of these technical systems that belong to the infrastructure of the Internet, which I call the 'technical apparatuses,' that are subsequently used on a Web-level to order content and users nationally. The software devices such as search engines and platforms that order content and users in distinct Web spaces are what I call 'technical arrangements.'

The term technical arrangement is used to call attention to how Web spaces are computationally governed. In *Seeing Like a State* (1998), political scientist James Scott follows a similar approach to examine how central modern governments attempted to force legibility on society. 'Seeing' like a state is a Foucauldian view on societal power structures whereby governmental institutions make their territory and people legible – organize and map – in order to govern most effectively. Scott examines the state's attempt to make society legible by arranging its territory and population, thus simplifying state functions by studying governmental models of thinking. One example of a governmental scheme Scott addresses is the permanent last name. The permanent last name helps central government keep track of their subjects. On the Web software devices arrange and organize Web territory and population to govern the Web space most effectively. The contribution of this study is a medium-specific approach to the study of national Web, which is a way to start thinking of the Webs as shaped and organized by technical arrangements. In the next chapter, the Webs as media of location from a technical perspective are discussed. Which technical indicators can be used to demarcate the territory of technical arrangements as well as study the national Webs?

3. The Webs as Media of Location

Help / FAQ / Content filters

What are content filters?

Flickr is a global community made up of many different kinds of people. What's OK in your back yard may not be OK in theirs. Each one of us bears the responsibility of categorizing our own content within this landscape. So, we've introduced some filters to help everyone try to get along.

There are 2 types of filters that you need to use for your content.

1. Safety Level

- Safe - Content suitable for a global, public audience
- Moderate - If you're not sure whether your content is suitable for a global, public audience but you think that it doesn't need to be restricted per se, this category is for you
- Restricted - This is content you probably wouldn't show to your mum, and definitely shouldn't be seen by kids

2. Content Type

- Photos / Videos
- Illustration/Art / Animation/CGI or other non-photographic images, or
- Screencasts / Screenshots - [what's a screenshot](#)

Figure 9. Flickr FAQ Content filters, 2009

On June 12, 2007, when the localized language version of the photo-sharing site came into existence, Flickr implemented a filtering system for potentially controversial photos (figure 9). Users registered with Yahoo! in Germany, Singapore, Hong Kong and Korea were prevented from seeing photos rated 'moderate' or 'restricted', thus only being able to access 'safe' photos. Flickr justifies this on the basis of local Terms of Service; these users are not able to turn this so-called SafeSearch off. Restricted photos are presented to the user as noise (figure 10).

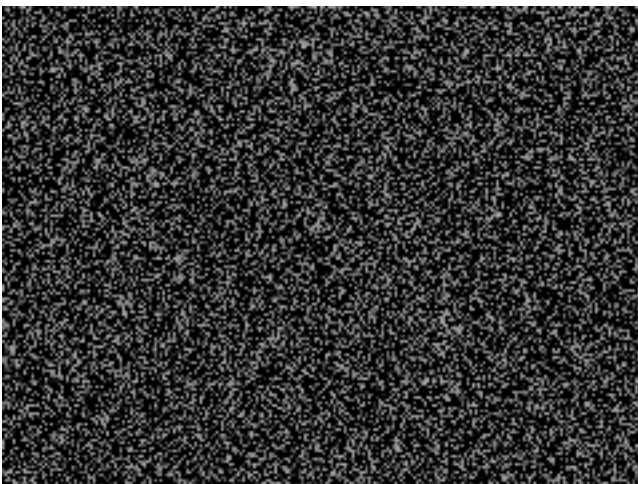


Figure 10. Blocked photos by Flickr

As with the YouTube story in the introduction where users were confused and discussed why they could not view certain videos, Flickr users discussed the nationalization of a presumably cyberspace service. Remarkably, however, the users complaining were mostly German. Users from Singapore, Hong Kong and Korea were noticeably absent.

In a selection of “[Official Topic] Filters,” an official Flickr Help Forum, “myfear”, an allegedly German user, just discovers the new Flickr content filters. “Loupiote (Old Skool)”, allegedly a hacker, helps “myfear” test to what extent the Flickr filters block him from viewing moderate photos. After ascertaining that moderate photos are not visible to the German user, the discussion moves on to the technical apparatus behind these content filters and ways to circumvent it (Figure 11).

myfear pro says:

Ok. Thats weird:
Can someone of the staff please comment on the new filter addition for german users?

Note: If your Yahoo! ID is based in Singapore, Germany, Hong Kong or Korea you will only be able to view safe content based on your local Terms of Service so won't be able to turn SafeSearch off.

Where does this come from? How to handle this in the future? I definitively do not agree with this behaviour!

what about the question from **loupiote (Old Skool)**:
Also, the FAQ says so won't be able to turn SafeSearch off. does that mean this restriction only applies to "Searches"? or does it applies to accessing any flickr photo-page accessed through users photostreams, sets, groups, etc?
Posted 21 months ago. ([permalink](#))

loupiote (Old Skool) pro says:

well, if you have a german account, it's easy to check.

try accessing any "moderate" photo page. here is an harmless "moderate" [artistic nude](#). can you access it?
Posted 21 months ago. ([permalink](#))

myfear pro says:

This photo is unavailable to you.

Great :) And now? Quit flickr ... the first time I am thinking about this ...
Posted 21 months ago. ([permalink](#))


loupiote (Old Skool) pro says:


this photo is moderate and harmless, as you can see by following [its deep link](#) (deep links are never censored).


but of course you need access to the photo page or to the thumbnail in order to know the deep link...
Posted 21 months ago. ([permalink](#))
loupiote (Old Skool) edited this topic 21 months ago.


loupiote (Old Skool) pro says:

and can you see the moderate photos (including this one) by accessing with [this guest pass](#)?
Posted 21 months ago. ([permalink](#))

 **myfear pro** says:
loupiote (Old Skool) I can see the tumbs in the sets overview. but if i click a single image the same happens ...
 This photo is unavailable to you.
 *p***
 Posted 21 months ago. ([permalink](#))

 **myfear pro** says:
 Ok... Any idea, from where flickr takes the account location information?
 Posted 21 months ago. ([permalink](#))

 **myfear pro** says:
 Great .. germany is like china ... blocking is even more effective .. there is no flickr-add-in that changes servernames and everything works fine .. thanks ...
 Posted 21 months ago. ([permalink](#))

 **loupiote (Old Skool) pro** says:
 well, at least if you can see the thumbnails, not all is lost.
 you can obtain the deep link by changing "_t" into "_d" in the deep link of the thumbnail, which can be obtained (on windows) by right clicking on the thumbnail and selecting "property".
 but this is really a hacker's way, and i wouldn't want to do that all the times. plus you don't have access to all the comments, tags, groups, sets etc attached to the photo.
 Posted 21 months ago. ([permalink](#))



 **loupiote (Old Skool) pro** says:
 Ok... Any idea, from where flickr takes the account location information?
 i guess, from your yahoo account.
 Posted 21 months ago. ([permalink](#))

Figure 11. [Official Topic] Filters, Flickr 2008

Yahoo! IDs based in Germany are not able to view restricted content due to local Terms of Service, imposed by German legislation (figure 12). After long discussions about the German settings, Flickr adjusted them slightly. At the time of writing, German users can view both 'safe' and 'moderate' content. 'Restricted' content, however, is still blocked.



swisskiltbear pro says:

Noluck

Yeah, legal requirements are different in every country. Germany is a good example. the "child protection" laws are so restrictive that it's almost impossible to host an "adult" .de domain. The age verification process is very complicated with a proposed "smart card login" (the same smart card that would have to be used in cigarette vending machines). So consequently, most german "adult" sites, have now moved to the Netherlands or other more liberal countries and are using .com or .info domains rather than .de.


Here in Switzerland, if I for instance sent a guest pass link for a set with restricted content to someone I personally know to be of legal age, I would be ok. If that person then would pass that link on to minors, it would be THEIR responsibility, not mine.

So yes, the matter of providing access to any restricted content directly from a site is very complex and I think for that purpose the **filters** work well. Just not for guest passes.

I am assuming though that it would be legally possible even with US law to amend the TOS so that responsibility for guest passes (which aren't readily available to those who don't specifically receive the email or link) is shifted to the issuer of the guest pass rather than remaining with Flickr.

Figure 12. Legal requirements, [Official Topic] Filters, Flickr 2008

Unlike YouTube, ABC, and most search engines, the Web arrangement Flickr bases its national filter on the Yahoo! email addresses serving as login for Flickr. If an email address ends with .de, the user is considered to be German. After the initial confusion, users quickly found out how Flickr restricted access to content and developed circumvention strategies. A different Yahoo! ID from a non-restricted domain like .com, is the most popular one (figure 13).



< insert fancy new name here > pro says:

solution for german users to avoid censorship:

- create a new u.s. based yahoo id. e.my.yahoo.com/config/my_init?_im=drawbridge&intl=us...
- transfer you flickr account to that new yahoo id. www.flickr.com/account/transfer/
- switch of safesearch. www.flickr.com/search/advanced/

if you want to renew your account with your german credit card.

- reverse steps
- pay
- transfer your flickr account to your u.s. based yahoo id again

it worked for me (at least part 1) and i really don't know if i pay again, if my pro account expires. thanks flickr, what a great move!

Figure 13. Censorship circumvention strategy, Flickr 2007

Assuming that Web content is ordered in a more and more national way, the various models arranging the Webs as such become an object of study. Different kinds of national thinking about the Web are made clear via attached technical apparatuses, such as the Domain Name System or IP-addresses. The Internet, and specifically the Web as organized along national lines may be studied in a number of ways. In this section locative technical indicators that can be used to demarcate, to a

certain extent configure and study the Webs are discussed, including top-level domains and IP-addresses. The national Web may be considered in parallel and in contrast with other technically defined Web territories, such as the previously mentioned TCP/IP model as the technical infrastructure for the cyberspace territory. The Spheres,' like the blogosphere, are also Web territories, and these are defined by ordering devices on the Web (Rogers 2007).⁸ The blogosphere, which reminds of Habermas' notion of the public sphere (Rheingold 2007), is no doubt best known, but it may be argued that there are other delineated Web spaces constructed by search engines and other ordering devices. The social bookmarking sphere e.g. can be considered as arranged by Web device Del.icio.us. Spheres are technical arrangements, but they are not necessarily national Webs. Thinking in terms of spheres is a device-centric approach to the study of Web spaces constructed by engines. They might include national Webs such as the study of the Dutch Web sphere constructed by search engine Google.nl.

The national Web, at least initially, is thought of in terms of 'locative media,' or, the media of location. Locative media are often discussed as user-generated cartographic information produced with location-aware mobile devices, such as receivers for global positioning satellites. The scope of locative media, "as opposed to the World Wide Web [...] is spatially localized, and centered on the individual user; a collaborative cartography of space and mind, places and the connections between them" (Tuters and Varnelis, 2006). As opposed to these location-aware mobile devices, this study focuses on the locative semi-aware technical spaces of the Web.

The technical approach distinguishes between 'grounding' and 'digitally grounding.' The first notion refers to the geographical 'grounding' of the Web, like in "the revenge of geography" (Rogers 2007: 1). The second notion 'digital grounding' refers the way society is embedded in the Web and how social trends can be distilled from the Web. For example, "Repurposing the WikiScanner" is a research project by Web researchers Erik Borra and Michael Stevenson that both 'grounds' a specific Web area and aims to 'digitally ground' social activity in that space at the same time. They propose to use the WikiScanner⁹ "to locate the production of Wikipedia knowledge within specific geographical and institutional borders. Rather than focus on acts of concealment or subversion - the individual deletion or addition," they try to unravel "the 'local' dimension of Wikipedia's collaborative authorship" (Borra and Stevenson 2008). In other words, they propose a research methodology to ground Wikipedia by mapping users to their geographical or institutional location and secondly analyze the correlation between geographical or institutional location and edit activity. Whereas 'grounding' is a statist approach to the Web and aims to localize and bring often-ambiguous geographical borders into the presumed virtual realm, 'digital grounding' is a national approach and aims to provide insights in the complexity of social interaction online. National Web research deals both with the localization of Web content and its users, and with the distillation of social trends in geographically or linguistically defined areas on the Web.

⁸ Strictly speaking 'ordering devices' such as search engines are devices that aggregate, order and serve third party content. 'Arrangements,' however, might also include services or platforms ordering and serving second party content, such as YouTube, Wikipedia and Facebook. Technically speaking these latter do not create 'spheres.'

⁹ The WikiScanner is a tool designed by Virgil Griffith. It 'de-anonymizes' edits on Wikipedia by linking IP-addresses from anonymous users to the organizations and institutions where the edits were made. The tool is mostly used for 'scandal research' (Griffith 2009).

Focusing on national Webs, it must be underlined that geo-locative technical elements are only one possible indicator for thinking about the Web from a locative-technical point of view. An indicator such as language, which can also be technically identified, is equally important for Web demarcation purposes. Although language does not necessarily coincide with national regions, it can be used as an indicator to demarcate language regions on the Web. Mobile phone networks or 'mobile Webs' use specific technical arrangements to geo-locate content and users. Geo-encryption is such a mobile Web technical arrangement, encoding streams of data in such ways that they can only be understood in a specific location. The output from a GPS device is used to unscramble the data, converting your location into your password (Lilley et al. 2006). Although not necessarily national, this technology enables thinking about the location of users and content in a locative-technical manner. These ideas about the Web distinct from the Web as a universal or placeless cyberspace focus on models arranging the Webs as a media of location, rooting and determining specific places.

Thinking about the Internet as cyberspace, Web technology does not have geography 'built in' like Global Positioning System technology does. It can, however, be arranged to become geographical. From a technical national Web approach, Internet technology has always had geography built in; it only needed to be configured as such. For example, YouTube's 'This video is not available in your country,' where IP-addresses are arranged to identify the geographical location of computers and their users. In order to extract the national from the Web, locative and natively digital elements that are embedded in the technical apparatuses of the medium are to be identified and combined. Hereafter four technical ways to think about the Webs as media of location in a 'grounding' manner are discussed.

Cybergeography

When thinking in terms of cyberspace, the technical reality of the infrastructure is subordinate to the notion of a placeless space. But in cybergeography, cyberspace is grounded in geography. The fiber-optic cable network allows computers to access the Internet wherever whenever, thus creating a distributed network distinct from traditional (de)centralized mass media (figure 14). When looking at the distribution of cables per country, it could be argued that cyberspace all along has had specific geographies built into its infrastructure (Rogers 2007: 2). Here concepts and technical reality collide. While cyberspace is all about a universal space, the fiber-optic cable network materially grounds it. The hardware infrastructure is structured geographically by means of its cables.

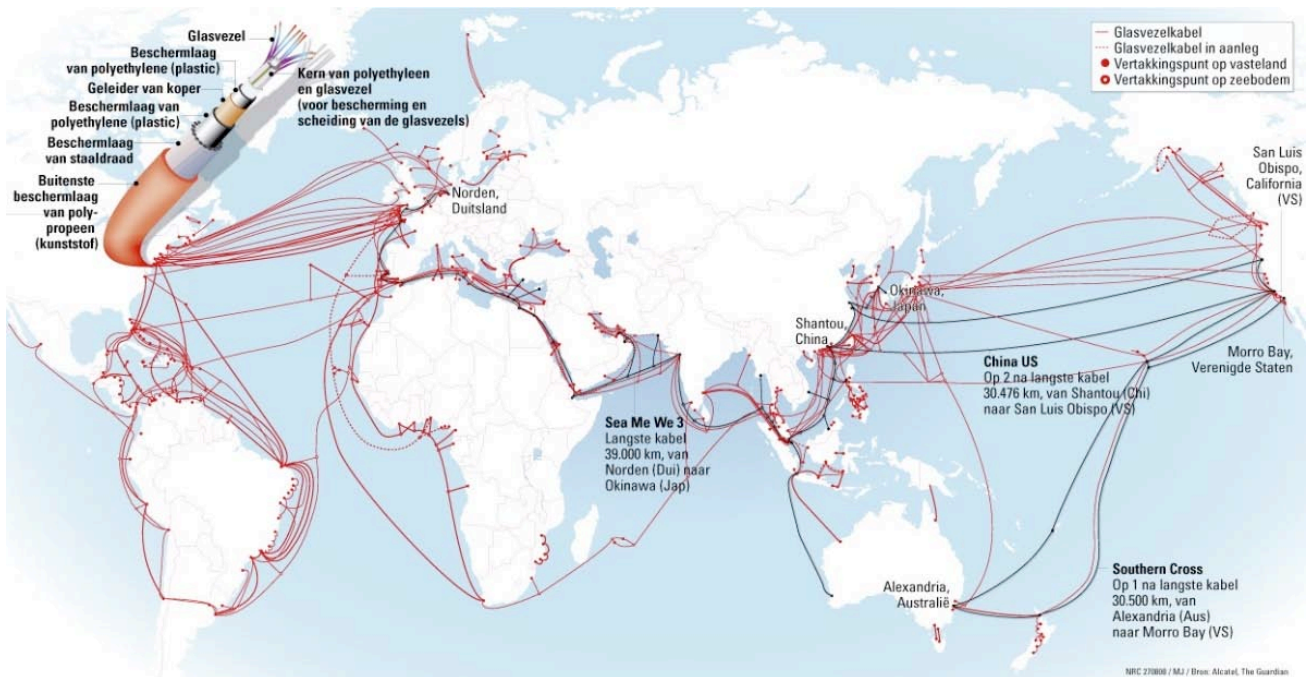


Figure 14. Cyberspace's fiber-optic cable network, NRC 2008

The Yahoo! Case: IP-to-geo

One important symbolic end of cyberspace in terms of locative technical indicators can be situated in the 2000 Yahoo! case. Marc Nobel, a French Jew who spent his life actively fighting neo-Nazism, came across a Nazi memorabilia auction site on Yahoo.com. Since it is illegal in France to traffic Nazi goods, he decided to sue Yahoo! on behalf of the International League against Racism and Anti-Semitism and others for offering this content on French territory. Yahoo! was summoned by the Paris High Court. This meant the start of reterritorialization of cyberspace, described in detail in *Who Controls the Internet? Illusions of a Borderless World* By Jack Goldsmith and Tim Wu (2006: 1-10). This Yahoo! case made Web epistemologist Richard Rogers proclaim the "death of cyberspace." Since the beginning of 2000, "you are sent home by default" (Rogers 2007: 1).

Yahoo vice president Heather Killen, initially argued: "We have many countries and many laws and just one Internet" (Tessler, 2000). Prior to 2000, the architecture of the Internet was not assumed to be built with geography in mind. Internet Protocol (IP) addresses, Domain Name System (DNS), nor e-mail addresses were conceived to geographically locate computers or Internet content. Almost half a year later, Cyril Hourri contacted the plaintiff's lawyer explaining that he had developed technology that could identify and screen content on the basis of its geographical source. At first, the geographical location of the servers offering the content was focused on; later, the key issue became Yahoo!'s ability to filter users by geography. The technical experts hired by Yahoo! developed IP-to-geo handling, which implies that the individual computer's IP-address can be cross-referenced with a set of IP-ranges assigned to a particular country. With IP-to-geo many countries allow for many Internets as different content can be served to IP-addresses from different countries. Within a universal cyberspace, it became possible to order and serve content according to local law, relevance, commerce or other factors.

Yahoo!, however, did not ban the Nazi memorabilia from the French Web, but decided to remove the content altogether. Only when Yahoo started to serve China and made a deal with the Chinese government, this technology was used to geo-locate users. The technology is currently applied by most major search engines to customize content and advertisement based on geographical location.¹⁰ In France, Google.com by default redirects to Google.fr. Google about location as a relevance measure:

By default, we identify your approximate city location based on your computer's IP-address and use it to customize your search results. If you'd like Google to use a different location, you can sign in or create a Google Account and provide a city or street address. Your specific location will be used not only for customizing search results, but also to improve your experience in Google Maps and other Google products (Garb 2008).

While the Internet Protocol infrastructure is distributed, a number of locative technical arrangements on the Web demarcate and order the Web according to national lines. This geographical order was always inherent to the medium and became embedded in devices and Internet institutions using the Internet's technical infrastructure. This reterritorialization of cyberspace is a medium-specific construction: it is embedded in the medium itself and it can be studied from within. Protocol setups can be configured to order and serve the Web along state lines.

The Web's central sites can be named. The Yahoo! and Google engines receive more hits than any other Website and can therefore be considered as primary access points (Alexa Top sites 2009). The Web's territories are ruled by devices such as Google, Technorati, Yahoo! and Del.icio.us, some of which deliver results by location and are enabled by institutions like Internet Assigned Numbers Authority (IANA), Internet Corporation for Assigned Names and Numbers (ICANN) or Regional Internet Registries (RIRs), stratifying the distributed Internet infrastructure. These devices, also referred to as 'portals,' first index and then order and serve the Web. Studying how these Webs are ordered and served along national or linguistic lines¹¹ is a way to ground and demarcate Web spaces. In all previously discussed examples (ABC, YouTube and search engines)¹² IP-to-geo location technology is used to serve customized results based on the user's geographical region. IP-to-geo is a medium-specific technical concept assigning geographical location to IP-addresses. Besides IP-addresses, there are other means to technically define objects on the Web. In the following other locative technical indicators are discussed.

The Domain Name System

The Domain Name System (DNS) is a technical apparatus that can be used in at least two ways to geographically locate content and users from within the medium itself. The country code top-level domain (ccTLD) system and the Whois IP-address registration can be used as locative technical in-

¹⁰ It should be noted that although IP-to-geo is considered as the beginning of customization, currently there are various ways to customize as well, including personalization.

¹¹ For certain global services there are country or language-specific versions, e.g., for Google as well as Wikipedia. Others have not, such as Alexa, which makes it possible to empirically challenge the notion of a national Web.

¹² See introduction.

dicators. A substantive amount of Websites can be separated into the 245 ccTLDs covering as many countries and territories in the world. ccTLD refers to the country specific top-level extension behind the final dot in URLs and mail addresses. Besides ccTLDs there are a number of non-geographical TLDs, also known as generic top-level domains, including .com (business/commercial), .gov (governmental), .org (non-governmental/non-profit organization) and .edu (education). All country domain identifiers consist of two letters, and all two-letter top-level domains are country domains. The 245 country domains tally with United Nations-recognized countries, but include also some non-sovereign islands and territories. The IANA, managed by ICANN¹³ creates and maintains country domains. IANA does delegate administrative responsibility of top-level domains to regional authorities, responsible for managing second or third level domain, such as .gov.ps, .com.ps.

IANA not only supervises the distribution of domains, but also the allocation of globally unique IP-addresses. It delegates registration to Regional Internet Registries (RIRs) that correspond to five major regions in the world (figure 15). The RIRs have Whois databases,¹⁴ which contain registration details of IP-addresses and domains. Using the Whois databases, The Measurement Factory visualized the geographical distribution of IP-addresses across the RIRs (figure 16).

Regional Internet Registries	RIRs	Region	URL
African Network Information Center	AFRINIC	Africa	http://www.afrinic.net/
Asia Pacific Network Information Centre	APNIC	Asia and the Pacific Region	http://www.apnic.net/
American Registry for Internet Numbers	ARIN	North America	http://www.arin.net/
Latin American and Caribbean Internet Addresses Registry	LACNIC	Latin America and the Caribbean	http://www.lacnic.net/
RIPE Network Coordination Centre	RIPE NCC	Europe, the Middle East, and Central Asia	http://www.ripe.net/

Figure 15. Table of RIRs

¹³ The process of domain registration was initially described by 'RFC 920: Domain Requirements.' This documentation specified the requirements for establishing a domain in the ARPA-Internet as well as the DARPA research community. Jon Postel at the Information Sciences Institute DARPA dealt with the registration process and maintained the database. At the turn of the century the ICANN.org took over management of the top-level domains including .com, .net, and .org as well as oversight of the IANA.

¹⁴ Whois is a widely used query/response protocol for querying databases to determine the owner of a domain name or an IP-address. Among other things, Whois databases can be used to determine the registration and hosting location of a particular Website.

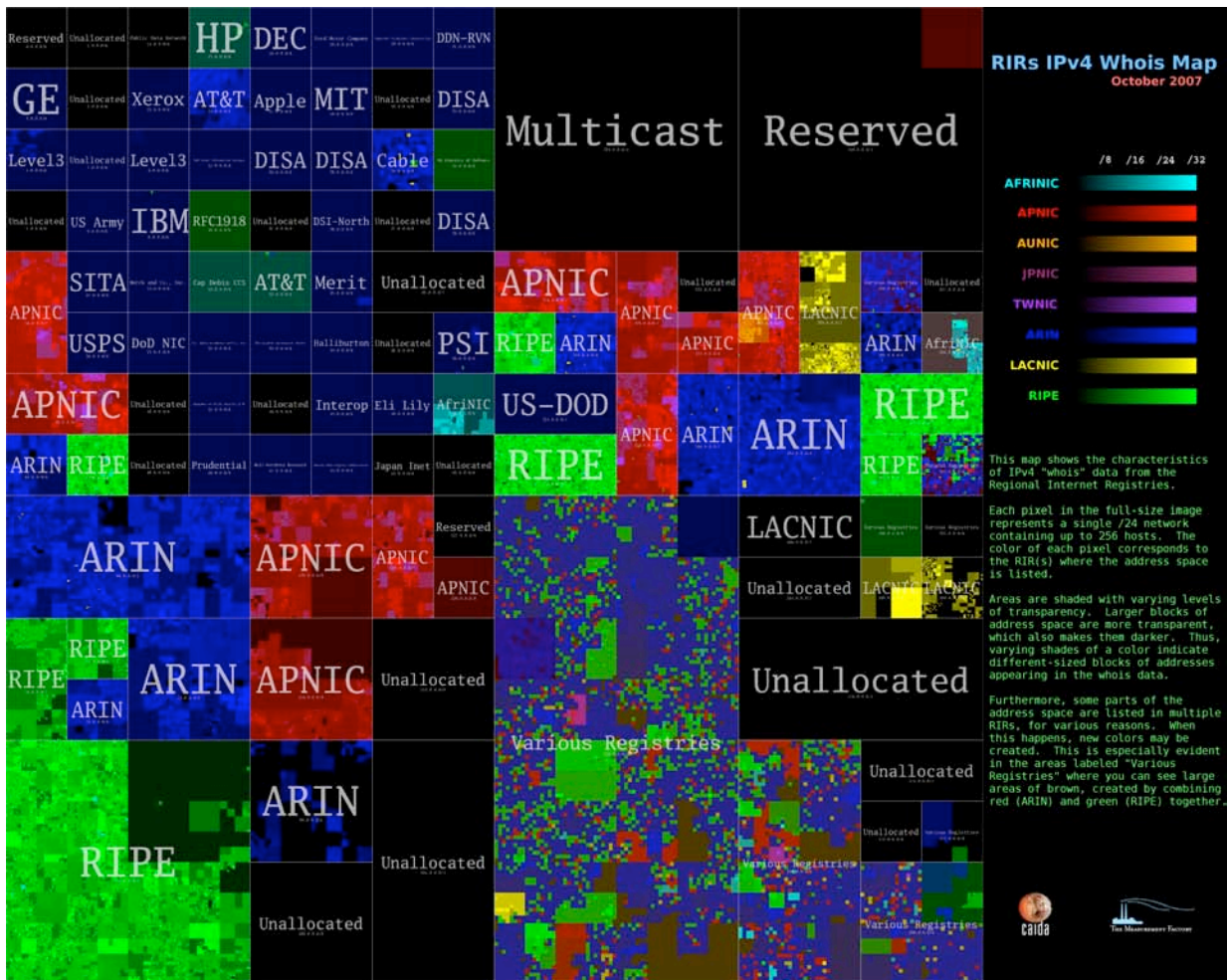


Figure 16. RIRs IPv4 Whois Map, Caida 2007

Locative technical indicators are not only useful for the central devices on the Web. Two projects demonstrate how these same indicators can be used to study the medium from within. The first is research in an ongoing 'Information Society in Palestine Project' and demonstrates how the Whois database can be used in Web research. A series of .ps Websites are queried for their registering and hosting location using the RIPE Whois database in order to locate and distill the national element from the Web. The resulting map (figure 17) shows a new geography of the Palestinian Web, the majority of .ps sites are registered within the Palestinian territories but hosted outside their borders (mainly in the United States).

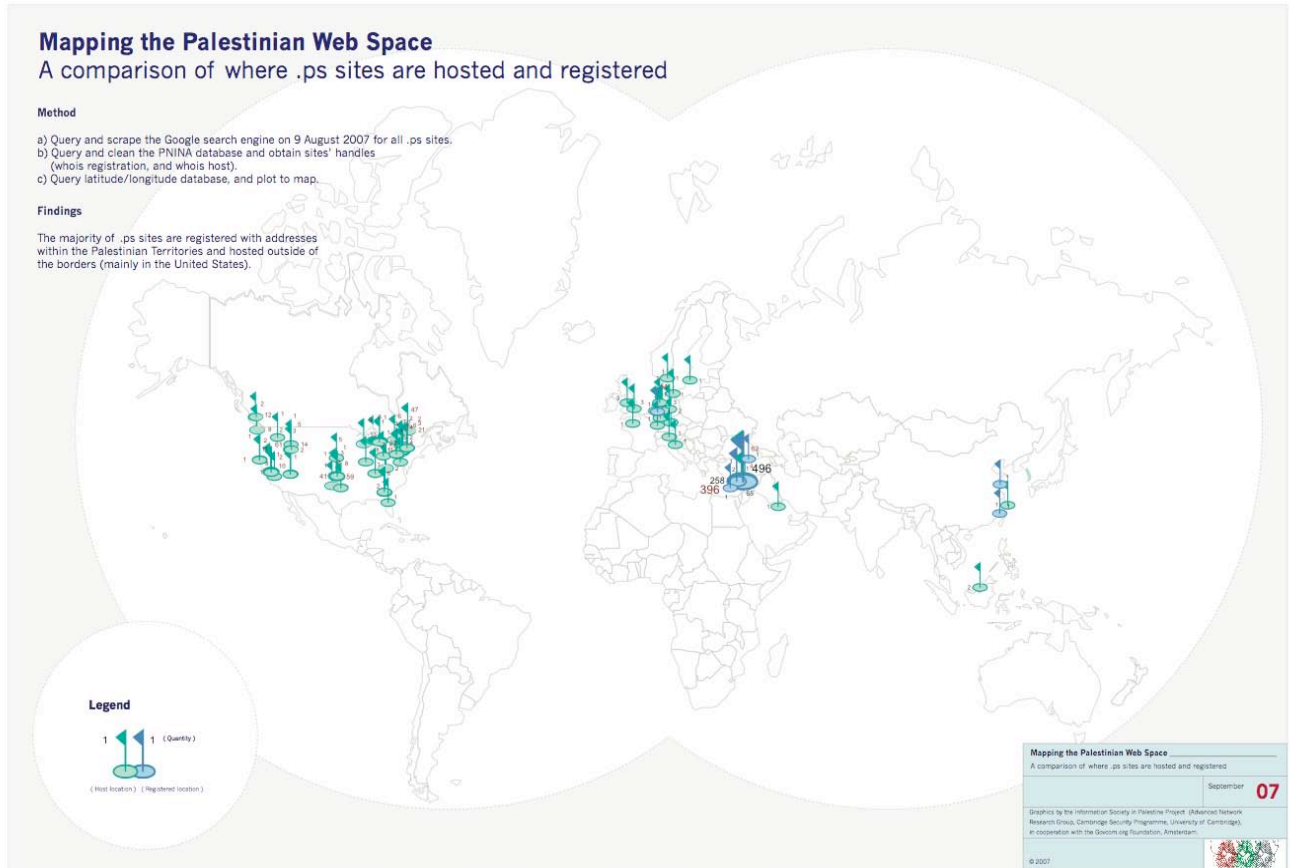


Figure 17. Mapping the Palestinian Web Space. A Comparison of where .ps sites are hosted and registered. Information Society in Palestine Project, 2007.

The second was inspired by a mapping project by research firm Byte Level that made a map of Web globalization, presenting ccTLDs in proportion to each country's or territory's population, with the exception of China and India, which were scaled down 30% to fit the layout (Figure 18). Offline data (country population) are used to inform us of the online.

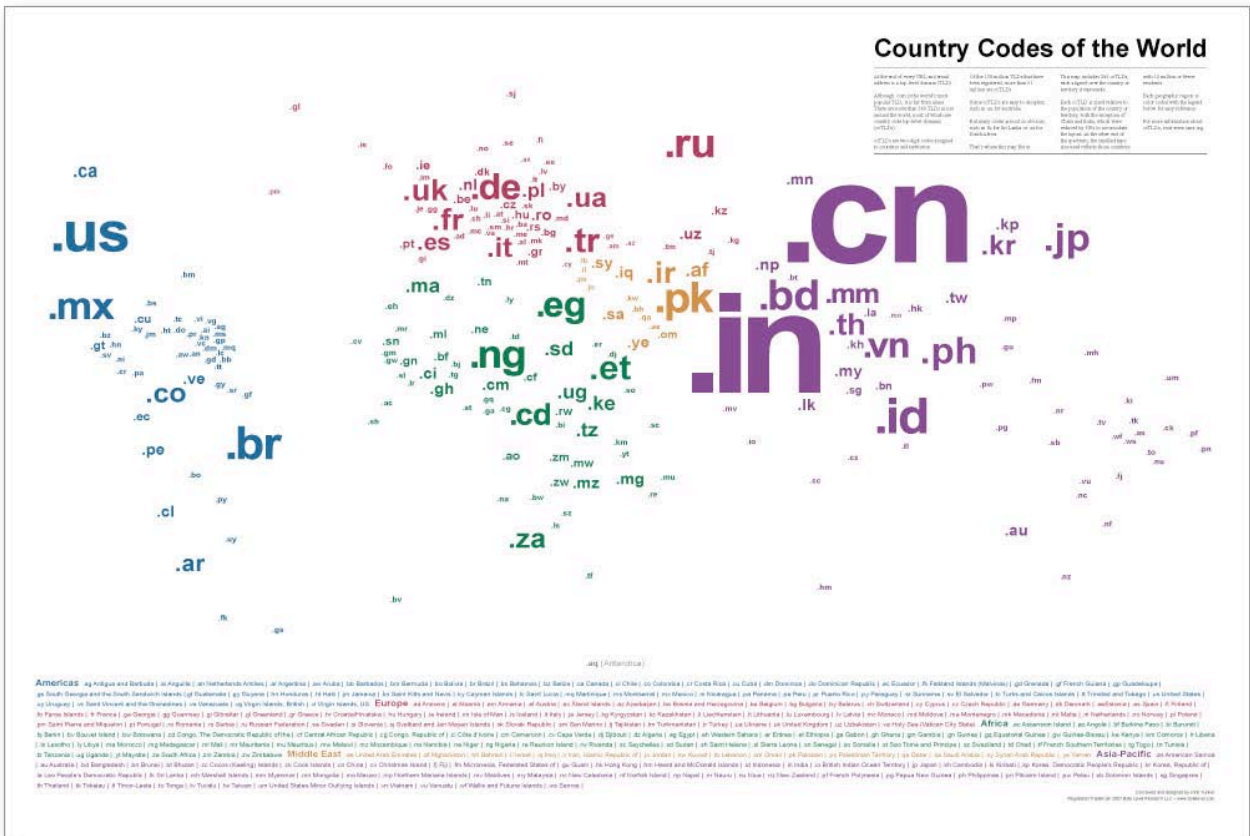


Figure 18. Country Codes of the World. Bytelevel Research, 2008

When studying the Web as a medium of location, however, the challenge is to technically locate and demarcate places of the Web. In a research project carried out by Erik Borra and myself for the Digital Methods Initiative, *The World According to Google* (figures 19 and 20), we used the locative technical point of view to study national Webs. The country domain Web map contains the estimated number of pages indexed by Google to size country domains. Google and the Domain Name System are thus used to study the national elements from a Web perspective. The method to make this map is: query Google for all country domains (e.g. "site:.nl," "site:.tv") and indicate them on a world map in proportion to other country domains. The map indicates the number of pages with a country domain instead of population. The global distribution of Web pages with country domains provides a radically different topology than the map by Bytelevel. As indexed by Google, Japanese, German, Chinese, Russian, and British domains have the largest number of pages online. This particular map tells us which country domains are the most actively used. Generic top-level domains are excluded from this map.

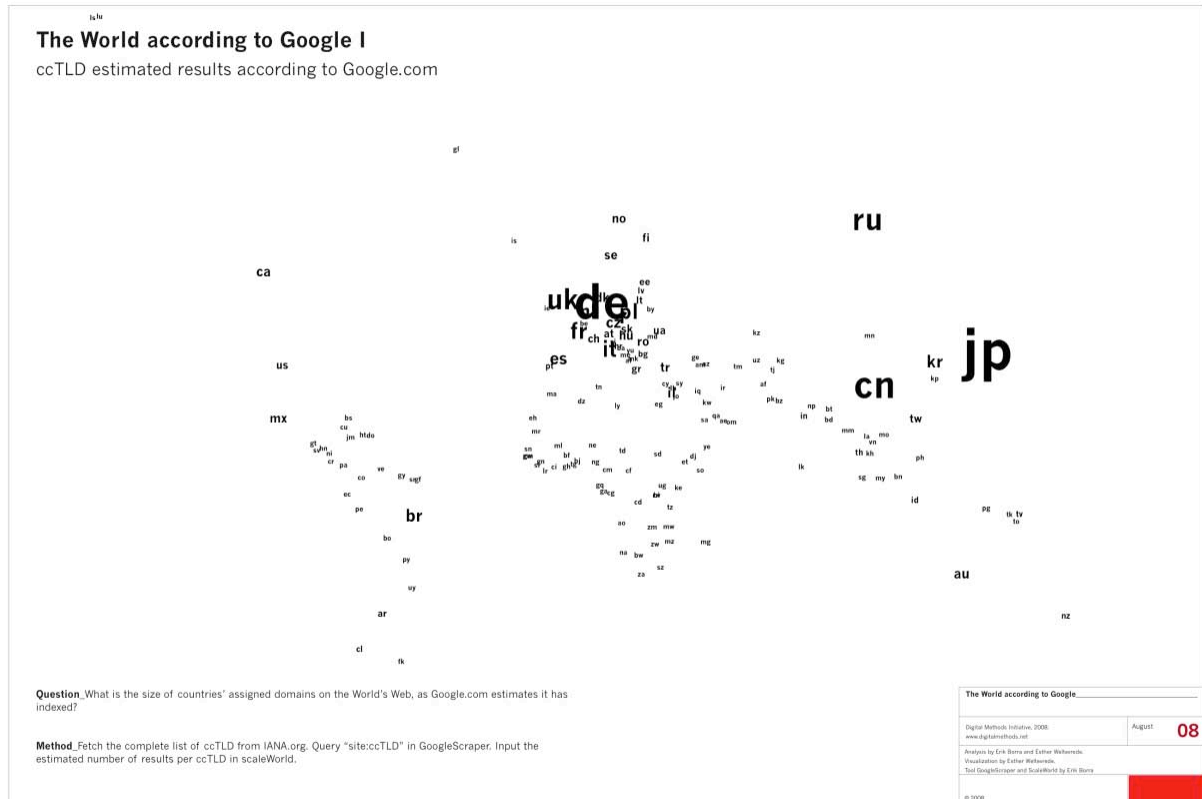


Figure 19. *The World according to Google I*, Digital Methods Initiative 2008

A country domain however, is not per definition related to the region it is assigned to. IANA arranges and delegates ccTLDs to designated local managers, who operate them according to a local policy so as to best serve the economic, cultural, linguistic, and legal circumstances of the country (Gomes 2006: 3). Every local authority, for instance, can decide over second and third level domains; these lower-level domains are not even documented by IANA. Local managers can also decide to what extent ccTLDs can be registered outside the geographical area. Appendix A shows whether ccTLDs allow foreign registration and includes a list of 'vanity ccTLDs', i.e. country domains licensed for worldwide commercial use because of their name. Tuvalu and the Federated States of Micronesia have a deal with VeriSign and FSM Telecommunications, for instance, to sell domain names using .tv and .fm TLDs respectively to television and radio stations.

A map of ccTLDs telling us something about countries' commitment to their assigned ccTLD, has been generated with data from Google's Region Search, the second in the series 'The World according to Google'. Region Search is a relatively novel Google Advanced Search feature, enabling to query from the Google's regional Web versions, which is basically a country version (i.e. Google.nl, Google.jp). Presumably, the term 'region' is chosen over 'country' because it includes territories as well. The map shows the most actively used country domains according to the Google regions. The map differs from the previous one since the estimated results for each country domain are now allocated to geographical region Google assigned the ccTLD to: query Google all country domains in the region they are assigned to (i.e. site:.nl in Region Search The Netherlands, site:.tv in

Region Search Tuvalu). Although the maps look quite similar, there are a number of smaller ccTLDs reduced in size in comparison to the previous map. Most notably is the disappearance of .cn from the map. The question arises, whether to what extent this might be explained by obfuscated results due to censorship practices in China (Villeneuve 2006). A list with the actual numbers can be found in (Appendix A).

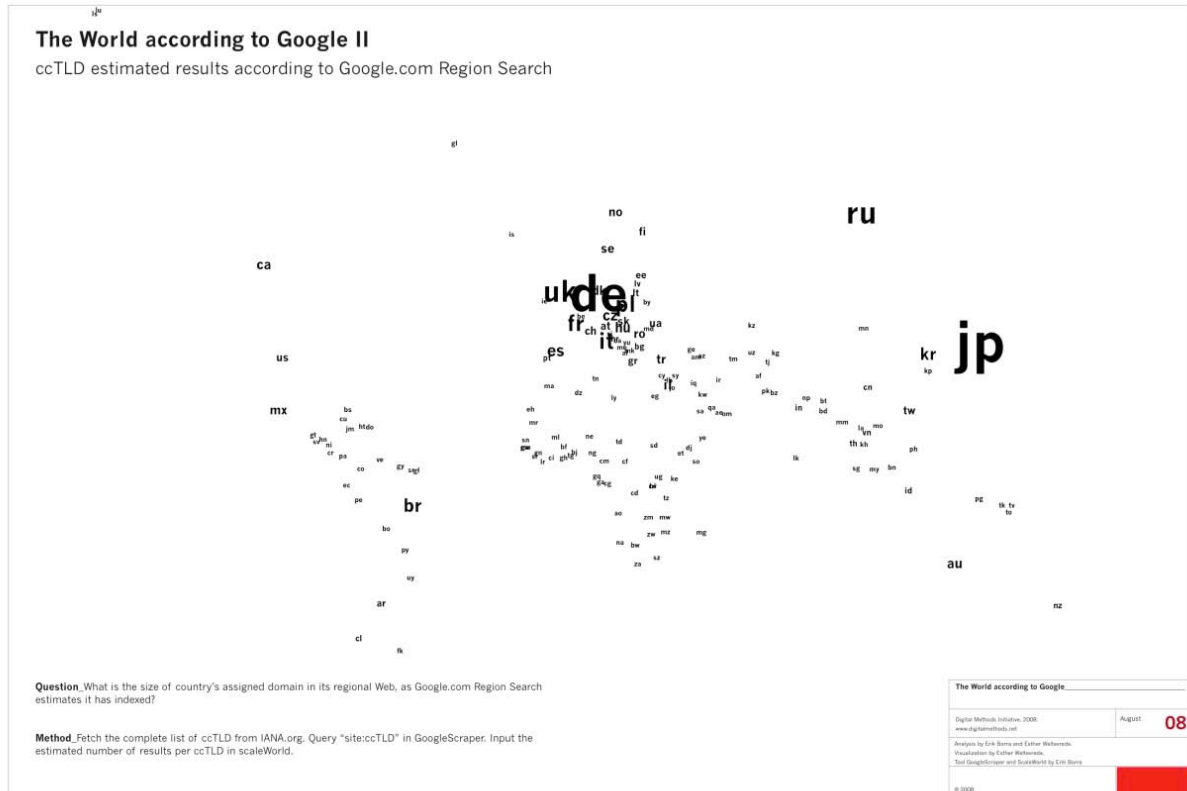


Figure 20. *The World according to Google II*, Digital Methods Initiative 2008

To show how a national Web's domain space is constructed, as well as to estimate the size of a particular national Web according to Google, one can use a similar method. By querying Google Region Search with a complete list of ccTLDs and gTLDs one can obtain any country or territory's topology.¹⁵ The United States and China, both economically important countries, have a domain topology dominated by .com, an extension referring to commercial sites (figures 21 and 22). More than half of the United States domain space is commercial. The Chinese Web has its country domain .cn as second most used extension, while in the United States .us is sixth, after .com, .org, .net, .edu and .gov. The United States also has a very elaborate range of second and third level country domains.¹⁶ Furthermore, when comparing the two domain topologies it is worth noting that both the United States and the Chinese Web have a diverse set of other country domains, which are used as vanity domains.

¹⁵ Note that these maps are generated using Google and are by no means a 'real' topology of a national Web domain. It is Google's estimate their indexation.

¹⁶ Cf. Wikipedia 2009 for a full list of .us country domains.

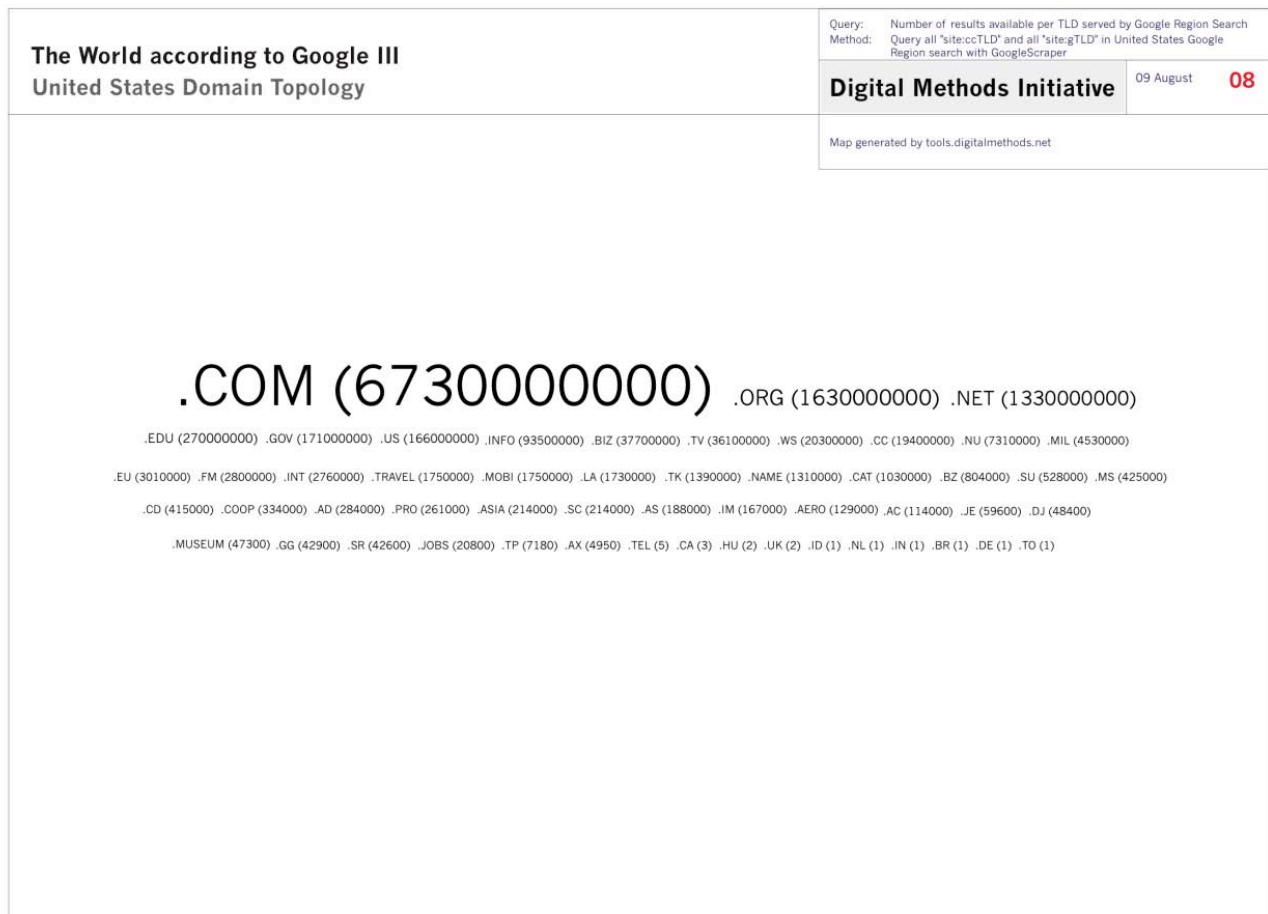


Figure 21. The World according to Google III, Digital Methods Initiative 2008

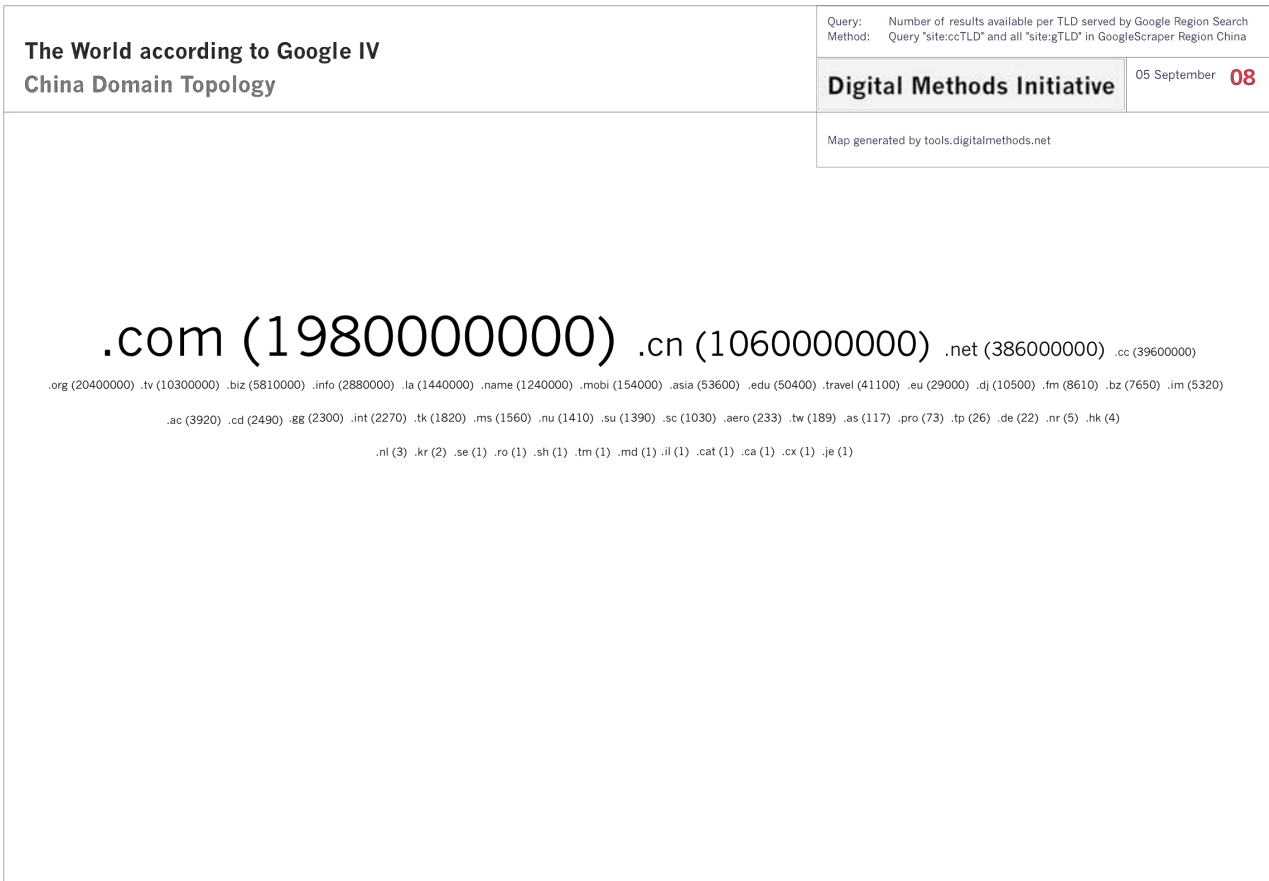


Figure 22. *The World according to Google IV, Digital Methods Initiative 2008*

Without studying the specific content of a national Web, the demarcation of the national Web's territory already seems to be able to tell us something about the socio-political situation. The Palestinian Territory Web rather consists of .org and .ps. than .com (figure 23). Furthermore, the domain space is less varied than the previous two. On the ground, the Palestinian Territory is not a sovereign state; while on the Web it has sovereignty, as Ben-David pointed out (Ben-David 2008). A Website owner choosing a .ps domain might therefore be considered to make a symbolic and patriotic statement.¹⁷ The predominance of .org in the Palestinian Web space is remarkably high in comparison to the previous maps.

¹⁷ A FAQ of The Palestinian National Internet Naming Authority (www.ps), "Why use .ps instead of ALL other domains?" contains the answer "Your country code Top-Level Domain (ccTLD) .ps communicates the Palestinian identity" (Palestine Info Society 2008).



Figure 23. *The World according to Google V, Digital Methods Initiative 2008*

Approaching national Webs in a medium-specific manner leans towards thinking about the Webs as media of location. Contrary to cyberspace approaches, this chapter has set out to demonstrate that the Web has geography built into its technology. Web arrangements use locative-technical indicators such as IP-addresses, top-level domains and Whois information to demarcate and configure the Web as national. These same locative technical indicators are then used in empirical Web research to first demarcate Web states and later analyze national Webs. Crucially, demarcating the borders of a national Web through top-level domains will draw different kinds of borders than through IP-addresses or Whois information. National Webs do not map one-to-one on state territories.

Next, Web archives will be analyzed and discussed as one particular Web arrangement. Here, too the national turn can be observed. Before analyzing two Web archiving projects, Web archiving as a practice is contextualized. The archives receive attention from a variety of disciplines, each with their own agenda for creating the ideal archive. In the theories of the archive, the historical and cultural relevance that is invested in the archive is approached by looking at the fields of humanities and social science, archival science, and lastly, it is discussed why the Web is a challenging archiving object when approached from a new media perspective.

Part 2: Web Archiving Principles and the National Turn

4. Theories of the Archives

The Web is increasingly recognized as cultural heritage. In the "Charter on the Preservation of the Digital Heritage," the United Nations Educational, Scientific and Cultural Organization (UNESCO) recognized the Web as a form of digital heritage (UNESCO 2003). Scientists from various fields seem to collectively agree on the cultural relevance of preserving the Web's digital information, including the humanities and social sciences. The archive is an important source for knowledge production, but only that what is saved in the archive can be studied later on. Whereas the aim of the first half of this study was to show that there is a national turn on the Web that can be approached in a medium-specific manner, the aim of the second half of this study is to find out why and how the Web archives look the way they do.

In this study the Web archives are approached as technical arrangements that select and order Web content to be included in an online-accessible archive. The first important archive on the Web was the Internet Archive, a cyber spatial initiative in the sense that it tries to index all Websites. Later a lot of Web archives have emerged with a national focus, exemplary of the turn to the national Web. In the following, first the history of archival principles and theories are discussed. These principles explicitly shape the object of preservation as they inform what is included in and excluded from a particular archive as well as how it is being ordered. Moving to the digital environment the question then is, what happens when archiving principles that have evolved from within their own field meet with the national turn on the Web? How does each type of archivist approach the object of collection and what consequences does it have?

Archiving for the Humanities and Social Sciences

As Jacques Derrida pointed out in *Archive Fever*, "Nothing is less reliable, nothing is less clear today than the word 'archive'" (1995: 90). Apart from within archival and library science, archives have been discussed considerably by other disciplines such as the humanities and social sciences. Jacques Derrida and Michel Foucault, the most influential theorists, initiated a discussion about the nature of archives. Their theory suggests converging interests with theorists, archivists, and librarians based on the purpose and future of historical and scholarly documents.

The standard dictionary definition of archives is two-fold. On the one hand it refers to "a place in which public records or historical documents are preserved," on the other "the material preserved—often used in plural," or "a repository or collection, especially of information" (Miriam-Webster 2009). Some authors state, that archives are for the humanities and social science disciplines what the laboratory are for the natural and physical sciences, because they both are sites of knowledge production (Osborne, 1999: 52; Withers, 2002: 304). Political theorist Irving Velody argues that archives stand as the backdrop to all academic research: "Appeals to ultimate truth, adequacy and plausibility in the work of the humanities and social sciences rest on archival presuppositions" (1998: 1).

In *The Archeology of Knowledge*, Foucault theorizes archives as a discursive knowledge system. He argues against archives as "the library of libraries;" or "the sum of all the texts that a culture has kept upon its person as documents attesting to its own past" (1972: 128-130). Archives are rather what he calls "the system of discursivity" establishing the possibility of what can be said

(1972:129). This theoretical concept of archives considers academic disciplines as systematic knowledge structures or discursive formations that define their own truth criteria. In his theory of “the history of ideas, or of thought, or of science, or of knowledge,” Foucault examines the continuity, as well as the discontinuities and ruptures of thinking within academic disciplines (1972). Discursive arrangements create the preconditions of what can be thought and said. All that ‘can be-said’ is secretly based on an ‘already-said.’ This “‘already-said’ is not merely a phrase that has already been spoken, or a text that has already been written, but a ‘never-said,’ an incorporeal discourse, a voice as silent as a breath, a writing that is merely the hollow of its own mark” (1972). It is supposed, therefore, that everything formulated in discourse is based on what is articulated previously.

The relationship between archives and political power, following a Foucauldian line of thinking, focuses on who files in archives and why. In “How Historians Play God,” cultural historian Robert Darnton states that, “in archives there lingers an assurance of concreteness, objectivity, recovery and wholeness” (Darnton 2002: 118). He argues that archives never consist of raw data, but are always constructed and therefore given direction for future use. As postcolonial scholar Edward Said argues, the ‘Orient’ is a textual construction of colonial culture. It has less to do with the Orient than with those who produce and preserve historical texts. More specifically, following Foucault’s idea of archives, he traces the construction of an Orient to Orientalism’s status as a discourse: “In a sense Orientalism was a library or archives of information commonly and, in some of its aspects, unanimously held. What bound the archives together were a family of ideas and a unifying set of values proven in various ways to be effective. These ideas explained the behavior of Orientals; they supplied Orientals with a mentality, a genealogy, an atmosphere; most important, they allowed Europeans to deal with and even to see Orientals as a phenomenon possessing regular characteristics.” (1979: 41-42.).

In *Archive Fever* Derrida, too, provides a way of thinking about archives that focuses on the technical arrangements of archives (1995). For various theorists in the humanities and social sciences, archives have always meant power. Derrida argues: “there is no political power without control of the archive, if not memory” (1995: 4). One of his most valuable contributions for this study is that the technical techniques and instruments used in the archiving process determine what can be archived, and that memory is therefore shaped by the technical methods of what he calls ‘archivization’ (Derrida 1995: 17). The recording devices inscribe traces in the archiving process that are included in archives. Derrida uses a counter-factual approach to explain what he means. He uses the example of psychoanalysis and how it would have completely changed the field’s history and development if Freud had had access to telephone, fax or computer (1995:16) as methods for sending and storing information shape the nature of the knowledge produced. Web archival technical methods shape what can be archived and studied. Derrida claims: “archivization produces as much as it records the event” (1995:17). The rules for inclusion and exclusion in Web archives are determined by for instance the crawlers’ code and the technical limitations and possibilities for locating Web content.

Archival Principles and Practices

The quest for knowledge rather than mere information is the crux of the study of archives and of the daily work of the archivist. All the key words applied to archival records—provenance, *respect des fonds*, context, evolution, inter-relationships, order—imply a sense of understanding, of 'knowledge,' rather than the merely efficient retrieval of names, dates, subjects, or whatever, all devoid of context, that is 'information' (undeniably useful as this might be for many purposes). Quite simply, archivists must transcend mere information, and mere information management, if they wish to search for, and lead others to seek, 'knowledge' and meaning among the records in their care.

- Cook, 1984

An important difference between archives and libraries is the notion of archives described as information generated as the 'by-product' of human activities, while libraries hold specifically authored information 'products' (Pearce-Moses 2005). The archival discipline is concerned with the circumstances (context) in which the information object existed and was used in order to serve as evidence and memory of historical facts and acts in the future. In what follows, archival principles and practices are introduced that demonstrate how archivists have a tradition of building information structures to preserve the context of records.¹⁸ Although archives have existed for thousands of years, most of the core concepts in archival paradigm were formulated between the mid-nineteenth and mid-twentieth centuries. A historical overview of archival collection principles is based on two articles describing the history of principles in the archival paradigm and their referenced literature. The first is by Canadian archival theorist Terry Cook, expert in the history of archives (Cook 1998). The second is by U.S. information scientist Anne Gilliland-Swetland, expert in electronic record management and fellow of the Society of American Archivists (SAA) (Gilliland-Swetland 2000).

In 1898 the Dutch archivists Samuel Muller, Johan Feith and Robert Fruin published their famous *Manual for the Arrangement and Description of Archives*. This influential publication summarizes the tradition of European archival theory and practice in a set of hundred archiving principles or rules. The first rule defines archives as "the whole of the written documents, drawings and printed matter, officially received or produced by an administrative body or one of its officials" (1898:13). Two important principles concern 'provenance' and 'original order.' The principles state that "the various archival collections placed in a depository must be kept carefully separate" and not mixed with other archives or placed into artificial arrangements (1898: 33). This is also known as *respect des fonds*, which entails that records should be grouped in archives according to the nature of the institution accumulating the records. The arrangement of archives "must be based on the original organization of the archival collection, which in the main corresponds to the organization of the administrative body that produced it" (1898: 52). In short, the Dutch authors advocate respect for the provenance, or 'birthplace,' of the records and the arrangement of the original record-keeping systems: the administrative context of state institutions.

¹⁸ This chapter focuses on archival principles and practices related to the selection and collection of records in the archive. For archival strategies on the long-term preservation of archival content, especially in the digital environment, including refreshing, emulation, replication and emulation see: Garrett et al. 1996; Rothenberg 1998 & 1999; Hedstrom et.al. 2003; Reagan, 2006

The Dutch Manual mainly deals with the arrangement and description of records in the archives. It hardly touches on 'appraisal,' which is another important concept in archival theory. Appraisal is defined as the process by which archivists identify and select materials of long-term value (Duff and Haworth 1993). In 1922 the then Deputy Keeper of the British Public Records Office, Hilary Jenkinson, published a second core book on archival theory and practice, entitled *Manual of Archive Administration* (1922). Jenkinson stressed the importance of the archives' function as 'impartial evidence' which can be defined as "the passive ability of documents and objects and their associated contexts to provide insight into the processes, activities, and events that led to their creation for legal, historical, archaeological, and other purposes" (Gilliland-Swetland 2000). If records are the by-products of administration and therefore evidence of acts and transactions, then the archivist's role is not to interfere with the collection; their role is to keep and not select archives. Instead, Jenkinson states that the original creator of the records best determines the value of the records and the selection to be archived. All together the principles of respect des fonds, provenance, appraisal and original order ensure that the intellectual integrity of collections of records is preserved and that individual records are contextualized. However, the consequence of Jenkinson's appraisal principle is that governmental institutions themselves are responsible for the selection of the records to be preserved in archives, resulting in archives reflecting the 'official' view of history.

Furthermore, Jenkinson introduced the concept of the 'archive group', with a different interpretation from respect des fonds. The archive group is more encompassing and might contain "fonds within fonds" (1922:102). They contain the entirety of records "from the work of an Administration which was an organic whole, complete in itself, capable of dealing independently, without any added or external authority, with every side of any business which could normally be presented to it" (1922:101). Archival theorist Jerry Cook recognizes the importance of Jenkinson's focus "on medieval and modern records, with their closed series, their stable and long-dead creators, and their status as inherited records from the past." Fluid administrative structures, he argues, "might create anomalies to challenge the archive group concept" (1998: 23). Although Jenkinson's views on appraisal are dated, since the stable nature of administrations as well as the fixed order of record arrangements has changed, his ideas on the evidential character of records remain prevalent in archival theory.

The United States began with professional archiving in the 1930s, facing an enormous backlog of governmental records. Facing the enormous amounts of records, archival theorist and State Archivist Margaret Cross Norton stated, in contrast to Jenkinson, that, "the emphasis of archives' work has shifted from preservation of records to selection of records for preservation" (Norton et al. 2003: 232). This led to the 'life cycle' concept, which means that records are first used and organized by their creators and after their operational use, a selection of valuable records is made by archivists (Cook 1998: 26). A pioneer in this life cycle appraisal theory was Theodore Schellenberg. He argued that records had primary and secondary values, primary value referring to its importance to their creator; secondary value referring to their importance to subsequent researchers. The archivist determines values after research and analysis. Instead of viewing all administrative records as 'archives,' for Schellenberg, 'archives' are only those records selected by the archivist for pres-

ervation from the larger whole, which he termed 'records.' Jenkinson's 'archive group' concept is replaced by Schellenberg's 'record group' (1956:146).

A fundamental change in the archival discourse is the 'societal approach' to the archives. The traditional institution or state-focused approach led to the perspective that archives should reflect the society the state serves. Hans Booms, an important thinker on the philosophical groundwork of appraisal, redefined collection procedures by stating that society must define the core values, which should then be represented in archival records. Booms wrote:

If there is indeed anything or anyone qualified to lend legitimacy to archival appraisal, it is society itself, and the public opinions it expresses - assuming, of course, that these are allowed to develop freely. The public and public opinion, as a constitutive element of modern society, sanctions public actions, essentially generates the socio-political process, and legitimizes political authority. Therefore, should not public opinion also legitimize archival appraisal? Could it also not provide the fundamental orientation for the process of archival appraisal? (Booms 1987: 104)

His contribution is that neither Schellenberg's expert archivists nor Jenkinson's state administrators, but society generates the relevance and values for the records in archives. The Canadian approach recognizes the intent behind the principles, which is "to link recorded information with the organic context of institutional (or personal) activity" (Cook 1998: 32). Appraisal is not focused on the records, but rather on governmental tasks, functions and activities that generate records. Developed from the early 1970s onward, the Canadian 'total archives' approach integrates the role of the archives as recorded evidence of transactions and the cultural role of societal memory. Canadian archivist Ian Wilson defines the 'total archives tradition as focusing more on the "records of governance" rather than on those of government (1995). Governance includes the interaction of "citizens with the state, the impact of the state on society, and the functions or activities of society itself, as much as it does the governing structures and their inward-facing bureaucrats" (Cook 1998: 34).

Helen Samuels, at the time of writing *Institute Archivist of MIT*, contributes to archival science in *Varsity Letters: Documenting Modern Colleges and Universities* by proposing a documentation strategy that links related personal records to complement institutional documents. She proposes a research-based approach to appraisal in order to locate relevant records in an interrelated information environment (across media types, as well as across institutions and persons) instead of focusing on isolated portions. This approach focuses on the relations between records based on activity instead of searching for values in the content of records. Important for this approach is the recognition that "analysis and planning must precede collecting" (Samuels 1992:15). These authors all advocate archives sanctioned in and reflexive of society instead of archives shaped by their document creators, often the state.

One of the developers of the 'total archives' concept at the National Archives of Canada, Hugh Taylor, advocates the societal approach to archiving. Taylor's essays and ideas exploring the nature of archives are influential. Canadian archival educator Tom Nesmith called Taylor's contribution "a rediscovery of provenance" (1993:4). Instead of considering provenance as a descriptive activity, this new understanding of provenance entailed creating a "historiography of the social"

(1993:4). Archivists began applying historical skills and methodologies to understand the social content, which generated the records. As Terry Cook noted in *Archivaria*, Taylor “was intent on constructing archives anew, imagining them as places where archivists connect their records with social issues, with new media and recording technologies, with the historical traditions of archives, with the earth’s ecological systems, and with the broader search for spiritual meaning.” This rediscovery of the information’s context focuses on the renewal of traditional archival principles such as “provenance, respect des fonds, context, evolution, interrelationships, order,” meant that archivists could move to the era of electronic records and networked communications without abandoning archival principles (Cook 1984).

In his 1966 article, “The Record Group Concept: A Case for Abandonment,” Australian archivist Peter Scott reinterpreted the concept provenance by focusing on description. He shifted from traditional ideas about archives as a static catalogue to a dynamic system of interrelations. He demonstrated that institutions creating records are generally not stratified along static hierarchical lines, but instead are a dynamic system of changing relations. He developed the Australian series system approach for describing multiple interrelationships between multiple records and creators. Although he developed this description system in the analogue 1960s, his insights are relevant for archivists of digital records. American archivist David Bearman, one of the most important thinkers in electronic archiving, asserts that the important point of these challenges to the traditional record is that “the boundaries of the document have given way to a creative authoring event in which user and system participate. Only the context in which these virtual documents are created can give us an understanding of their content” (1990: 11).

The Web as an Ephemeral Archive

The third theory of the archive discussed here approaches the Web as a database medium. One of the greatest oddities of online accessible Web archives, is that the archives become intertwined with the collection object. By making the archive available online, the distinction between the ‘live’ Web and the archived Web disappears. Social theorist Adrian Mackenzie argues that the intertwining of a ‘live’ Web and its archived information is a central feature of the database medium, such as the Web. The content databases’ structure plays a significant role in what he calls the “real-time and archive drives: “not only does the structure of the archives increasingly determine the coming into existence of its contents, these contents exist as real-time deferred” (Mackenzie 1997: 68). The real-time element only exists as a presentation on the screen. For Mackenzie, the centrality of the archive to cyberspace stems from the fact that existing or being in cyberspace is premised on a live connection to the archive (1997: 66).

The Web, resembling one vast, rapidly fluctuating archive is, unlike a traditional archive, being rebuilt every minute. Its sites can disappear within days, hours or seconds. Web content is revised and updated, often leaving no records of the previous alterations. Viewing the Web on the one hand as an archival medium and ephemeral medium on the other, the two notions seem to challenge each other. In “The Enduring Ephemeral, or the Future is a Memory” Wendy Chun unravels the notion of ephemerality as “the conflation of memory with storage” (2008: 1). The majority of digital media is memory and is placed as its ontology at all levels, from hardware to software, from

content to purpose. However, it is this conflation of memory and storage “that both underlies and undermines digital media’s archival promise” (2008: 1). Storing is static, while memory, such as RAM, is a process. This conflation is not based on some inherent technological feature, but is rather due to everyday usage and manner of speaking. Unpacking the theoretical implications of constantly disseminating and regenerating digital content, she argues that this conflation of ideas creates, rather than solves, “archival nightmares” (2008:1).

The Web, as a rapidly fluctuating archive is an object that is not simply ready to be archived. Archivists have to deal with the ephemeral nature of the database medium and discriminate to identify what is significant from a mass of data. Although the Web has its own specific challenges, the history of archival principles indicated archiving has always been about discriminating what is included and excluded from the archives. These principles call attention to who does the archiving, what are archives and why.

Unlike other well-known media, the Internet does not simply exist in a form suited to being archived, but rather is first formed as an object of study in the archiving, and it is formed differently depending on who the archiving, when, and for what purpose.

- Brügger, 2005

There is a lot at stake for the humanities and social sciences. As argued by Foucault, for these disciplines, the archives establish what can be said in discourse; they are sites of knowledge production. By stressing the importance of the ‘shape’ of the archive for the humanities and social sciences, attention is drawn to the material selected for the archive and how it is being archived. Derrida calls attention to the technical methods of archiving as shaping the nature of knowledge produced and recorded. In the Web environment archivization is inscribed in the technical methods and techniques. These methods include hardware-based as well as software-based tools and need to be built in series and be made compatible. In various stages of Web archiving tools, among others, execute the collection process, manipulate the archive format and make the archived collection accessible to humans.

The medium-specific approach of this study strives to critically examine the archives’ shape by the technical methods used. The endeavor is to find out why and how the Web archives look the way they do. In the following the first Web archive will be analyzed and discussed. The Internet Archive is an archive that aims to save the entire Web for posterity. It is the most extensive Web archive to date. How do the Internet Archivists approach their object of collection? What are their technical methods? And what kind of archive is the result?

5. Archiving cyberspace: the Internet Archive

“Only the exhaustive is interesting”

Ludwig Wittgenstein in Greetham 1996

The United States-based Internet Archive is the first and most exhaustive initiative in the field of Web archiving.¹⁹ Its founder Brewster Kahle dreamt of archiving the entire Web: “I usually work on projects from the you've-got-to-be-crazy stage,” Kahle says, “but eventually everyone ends up saying, ‘Of course’” (Reiss 1998). In another interview from around the time the Internet Archive was launched, Kahle says: “given that many of the world's greatest books, lyrics, images and other artworks pass through these networks, an Internet Archive would, some say, be nothing less than the sum of all human knowledge” (Kahle in Boyle 1997). The aim is to find out how and why the Internet Archive is shaped the way it is. The analysis of the interface of the Internet Archive and the archivists' writings is focused on the extent to which the period the Web archive was created as well as the archivists' approach to their object of collection shaped the archive.

Web archiving is the process of locating and collecting portions of the Web, preserving them in archives, but also making them accessible to researchers, historians, and the public. The archivist's approach is the long-term preservation of Web content in the archive, while the librarian's approach is making the collection accessible and searchable, which is invested in the search interface on top of the archive: the Wayback Machine. The reach of the archived Web as well as its searchability depends on certain technical limitations but also on the period in which it was conceived. In the following, the mindset of the early Internet Archivists is looked at in order to find out how it shaped what Web has actually been preserved and how it was made accessible.

Fortunately, the Internet Archive attentively self-preserved its Website in their archive. In 1997, the first archived page of the Internet Archive, accessed via the Wayback Machine, reads:

Internet Archive. Building a Library for the Future
Mission Statement

Internet Archive is collecting and storing public materials from the Internet such as the World Wide Web, Netnews, and downloadable software which have been donated by Alexa Internet. The Archive will provide historians, researchers, scholars, and others access to this vast collection of data (reaching ten terabytes), and ensure the longevity of this information. For more information about our philosophy and objectives, please read *Archiving the Net* by the Archive's founder, Brewster Kahle (Internet Archive 1997)

This initial mission statement already contained a number of larger topics, still dominating the Internet Archive project, as well as the Web archiving field. The Internet Archive's broad collection, larger than the Web, including cyberspace areas such as Netnews, intends to preserve the Web's content for analysis and use by historians and scholars for times to come. Following the narrative suggested by the Internet Archive, Brewster Kahle's article “Archiving the Internet” (1996) provides

¹⁹ The Internet Archive also includes a Moving Image, Live Music Archive, Audio and Text database, but these are excluded from the analysis. The focus in this study is on the Internet Archive's Web archive. When the 'Internet Archive' is mentioned it refers specifically to the Internet Archive's Web archive.

a more detailed description of the specificity of the Internet Archivist's dreams and thoughts about the collection's object and objective. Some of these dreams and thoughts are still relevant in the current Web archiving context; others have changed somewhat due to technical developments, or have been abandoned all together.

Internet Archivists' dreams from the beginning of Web archiving had to come to terms with the then existing medium's concepts, such as the trustworthiness and quality of Web information, predominant questions in the nineties. The initial framing therefore was institutional, to give this project weight. The first special collection, regarding the 1996 presidential elections in the USA, was carried out in collaboration with the US Library of Congress. Kahle compares his efforts to those of the ancient librarians of the Library of Alexandria, which went up in flames in 389 AD. According to Kahle's estimate of bits and bytes, the 1997 Internet Archive was more than twice as large as Alexandria's papyrus's store. The images on the first archived site demonstrate this library of bits and bytes (figure 24), rewarding the Internet Archive the status of a traditional library, but also pointing out that the loss of content with cultural relevance should be avoided. Today's cultural artifacts might be valued differently in future, especially by historians and scholars; "the history of early materials of each medium is one of loss and eventual partial reconstruction through fragments" (Kahle 1996). Information and library scientist Peter Lyman suggested in his paper "Archiving the World Wide Web" (2002) that the Web is the information resource of first resort for millions of readers. The new object of collection, Web material, is unprecedented in spreading the popular voice of millions, which were, until the Internet Archive, not saved for future generations. In the following few sections other period based considerations will be stipulated.



Figure 24. Internet Archive is creating a library of bits and bytes, Internet Archive 1997

Cyber Spatial Concerns

According to Kahle, Web archives might help to deal with some common infrastructural complaints, dominating the 1996 Web. These complaints reflect issues concerning the medium as cyberspace. Kahle highlights some of the dominant thoughts (1996):

- o Internet seems unreliable: "Document not found"
- o Information lacks context: "Where am I? Can I trust this information?"
- o Navigation: "Where should I go next?"

Brewster Kahle, via the Internet Archive, as well as the Web Information Company, Alexa Internet, tried to make the cyberspace browsing experience smoother. One of his core objectives was solving the '404 Document Not Found' error messages problem. The Internet Archive and Alexa try to do the same: "Alexa promises to banish '404 not found' messages for its members by retrieving stale pages from the Archive" (TBTF 1997).

'Can I trust this information?' was a discussion from the 1990's focusing on the quality of information on the Web in general; recently, however, the focus was changed to Wikipedia in particular (Stvilia et al, 2005; McGuinness et al, 2006). Others have pointed out that the Web has its own mechanisms to determine the reputation and value of information (Rogers, 2004; Sunstein, 2006). The rise of ordering devices as 'portals' to the Web both assign value to content as well as provide context for navigation. Think for example how Google 'ranks' the value of content with their powerful PageRank algorithm, essentially by counting and measuring hyperlinks to a particular Webpage.

Instead of browsing through cyberspace, current Web visions are dominated by searching. In current search practice, 'Where am I?' and 'Where should I go next?' are issues of a different nature. Instead of following hypertext paths from one Webpage to the next, research has shown that current navigation focuses more around the result pages of search engines. Expert users have developed search strategies using search engine result pages into a context reference (figure 25). Users with "Internet-related knowledge" have specific search strategies to quickly determine the relevance of information by using search engines (Hölscher and Strube 2000). As opposed to browsing, challenges in searching include choosing the most suitable search engine or service, query-formulation and the assessment of sources returned.

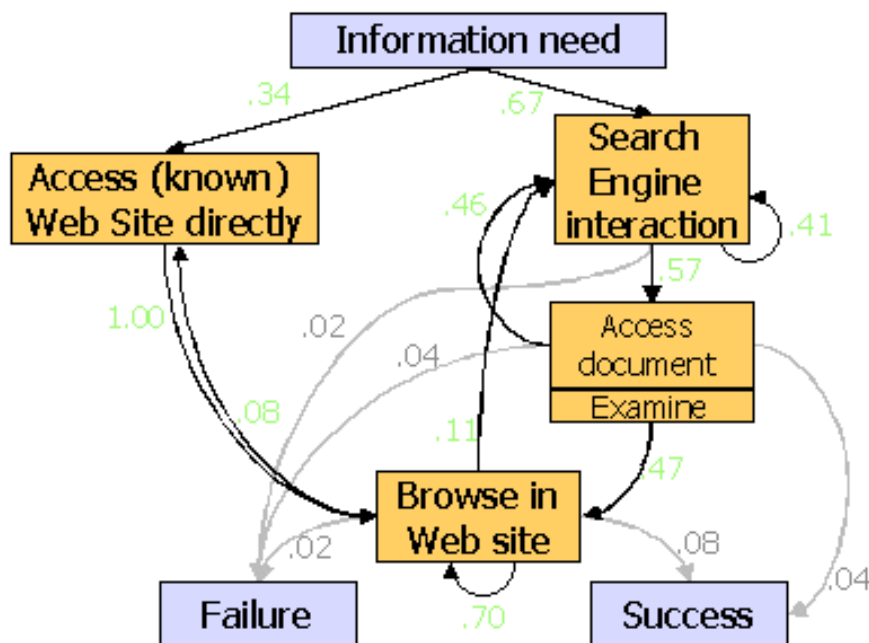


Figure 25. Expert user search strategy with a central position for the search engine, Hölscher and Strube 2000

Cyberspace in a Box

The early publications about the Internet Archive lack illustrations of cyberspace: There are no maps.²⁰ There is however an image of how the archived collection is stored. Just like cyber-geographical maps of cables, the hardware infrastructure of the Internet Archive is discussed. A traditional library requires an enormous building to store its entire collection, but Kahle imagines the Internet library to fit into a box. In the 1996 *Wired* article “Internet in a Box,” which is also referenced on the first Internet Archive Website archived, the box he imagines is “an ADIC 448 tape robot,” which “looks like a cross between a jukebox and an extra-tall dishwasher” (figure 26) (Reiss 1998). With a little compression, Kahle says, “the whole World Wide Web should fit pretty nicely onto one of these” (1998). Because of the dropping cost of data storage, capturing the information is not the most difficult thing to do, he argues. Initially tape jukeboxes combined with hard disk storage were chosen.



Figure 26. Tape Jukebox ADIC Scalar 448

In 2004 the Internet Archive introduced the Petabox storage system (figures 27 and 28). The largest part of the Internet Archive is stored on hundreds of slightly modified x86 servers. The computers run on the Linux operating system and each have 512Mb of memory and can hold just over 1 Terabyte of data (Internet Archive FAQ 2009).

²⁰ For an extensive collection of maps, see the *Atlas of Cyberspace* (Dodge and Kitchin 2008).



Figure 27. "The bits go here. A sample Internet Archive server rack, encompassing a petabyte of storage. A petabyte is 1000 terabytes, and a terabyte is 1000 gigabytes," Kirschenbaum 2007; Petabox 2009



Figure 28. The physical data center of the Internet Archive, Bibliotheca Alexandrina 2009

The Internet Archive, which has preserved the Web since 1996, was shaped by the then dominant thoughts about the Web as cyberspace. Navigating the Internet Archive through its interface, the Wayback Machine, is by searching a URL and browsing from page to page, instead of the now current practice of accessing the Web through keyword queries. Approaching the Internet, as one universal medium and a smooth browsing experience as challenge are both cyber spatial indicators.

The Internet Archive has emerged from the Web company Alexa and has been closely working together at the early stage of the archiving process. On the other hand, the Internet Archive

closely works together with traditional libraries such as the Library of Congress and aims to create the library of the Internet. The effort is to find out whether the Internet Archive is first informed by archival and library principles and practice or by what is technically possible. Put differently, do the Internet Archivists approach the archiving process as librarians, archivists or as Internet experts?

The Web Archive and the Digital Library

Web archives and digital libraries can be considered to be conceptually comparable, as they both store and make accessible digital contents. However, it is generally accepted that digital libraries focus on the accessibility of digital information whereas archives focus on long-term preservation (Gomes et al. 2006). In the "CNI White Paper on Networked Information Discovery and Retrieval," Clifford Lynch defines the digital library as an "electronic information access system that offers the user a coherent view of an organized, selected, and managed body of information" (1995). The effort of digital libraries can be summarized as striving for "Universal Access to all Knowledge" (Kahle 2004).

Web archives, on the other hand, focus on the long-term preservation of Web content and generally entail gathering large collections of records from the Web, without or with limited human intervention (Gomes et al. 2006). Open Archival Information System (OAIS) uses the term archive (ISO standard) when referring to an organization that intends to preserve information for access and use by a designated community (ICPSR 2007). Digital archiving can refer to an encompassing collection of digital and digitized documents, while the collection of Web archives usually refers to 'digitally born' material, i.e. material that is not digitized from an analogue original, but rather relates to a record that was created and exists only in a digital format (ICPSR 2007). Web archives should thus not be confused with either digital libraries, which collect and provide access to digital information, but may not commit to its long-term preservation, or digital archives, which do include long-term preservation but include imported digitized material in their collection.

The mission statement of the Internet Archive is "building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public" and therefore considers itself both an archive and a library (Internet Archive 2009). The Web archive for long-term preservation as well as a digital library that orders, structures and makes accessible the archived content are however addressed in technical terms. Challenges for archival and librarian issues are met with Internet expert means and resemble techniques used by digitally born Web devices such as search engines. The 'archive' part is invested in the backend of the Internet Archive and includes customized crawlers to fill the database with archived Web material. The 'library' part can be located in the front-end interface to the Internet Archive: the Wayback Machine. How these are made to work with and for each other is discussed next.

Heritrix, Metadata and the Wayback Machine

A crucial aspect of preserving Websites is metadata. The term metadata has different meanings. In library and archival science metadata frequently refers to cataloguing and forms of descriptive information, but it can also refer to information about the administration, preservation, use, and techni-

cal functionality of digital information resources (Gilliland-Swetland 1998). On the backend of the Internet Archive, the Web crawler Heritrix indexes and archives Websites for the archive. In the archiving process Heritrix collects metadata. The focus is on the metadata that is subsequently used for the front-end of the Internet Archive: the Wayback Machine. The question is, how are archival principles implemented in the technical? In other words, what type of (meta)data does the Internet Archive store in the backend and in what context are information objects placed in the front-end?

Internet Archive's open source Web crawler Heritrix, aka 'ia_archiver,' is continuously being redesigned to cater for new archival needs. Search engines also make use of crawlers to index the Web, which they subsequently make accessible via the search interface. The special crawler Heritrix is different from a search engine crawler because it does not only discover and index Webpages, but on top of that stores a copy on the local Internet Archive's servers. The crawler is sent out to collect the 'born digital' records, which are digital documents such as HTML, PDF and JPEG files. The order Web archivists impose on the Web's knowledge objects is invested in the library part of the Internet Archive: the Wayback Machine. This order, however, is limited and enabled by the metadata the Web crawler Heritrix collects.

The Crawler User Manual contains its settings to define the scope of crawling (2009). These settings define where a Website begins and ends. The crawler gathers metadata in the archiving process, which are important for the information design of the Wayback Machine. The metadata collected at each crawl are mostly based on technical crawl data (Library of Congress 2009). The metadata logs of the crawl contain a URL, crawl status (whether or not successful), a time-stamp and size of the downloaded document (Kahle 2007). The collected metadata can subsequently be used to structure and make accessible the born digital through the Internet Archive's Web interface.

Alexa engineers in cooperation with the Internet Archive, created the Internet Archive's Wayback Machine (Internet Archive About 2009). The FAQ "What is the Internet Archive Wayback Machine?" says, "The Internet Archive Wayback Machine is a service that allows people to visit archived versions of Websites. Visitors to the Wayback Machine can type in a URL, select a date range, and then begin surfing on an archived version of the Web" (Internet Archive FAQ 2009). The Wayback Machine, which is made with the cyber spatial browsing in mind, can only be queried by URL. The result page is a list of dates, ordered by year, of the URLs archived. The dates may include a '*', which indicates that the saved version is different from the previous one (figure 14). The Internet Archive assigns a URL within the site to the archived files in the format: `http://Web.archive.org/Web/%5BYear"http://Web.archive.org/Web/[Year_in yyyy][Month in mm][Day in dd][Time code in hh:mm:ss]/[Archived URL]`. Typically, the status bar from a Web browser will show the original URL in the footer. The archiving date assigned applies to the HTML file but not to image files linked therein. Therefore, the images that appear on the retrieved page may not have been archived on the same date as the HTML file. Likewise, if a Website is designed with 'frames,' the date assigned applies to the frameset as a whole, and not to the individual pages within each frame (Archive Legal 2009).

All links on archived pages refer to other pages in the Internet Archive if available. "Not every date for every site archived is 100% complete. When you are surfing an incomplete archived site the Wayback Machine will grab the closest available date to the one you are in for the links that

are missing. In the event that we do not have the link archived at all, the Wayback Machine will look for the link on the live Web and grab it if available” (Internet Archive FAQ 2009). The best way to be sure about the date of the archived file is by looking at the date code embedded in the archived URL.

Not all images or files are archived, however, which may lead to the presentation of only the skeleton of some retrieved pages. According to Kahle, some missing images are due to robot exclusions, but sometimes it is due to changes in the crawler. “From 1996-1998, the crawler crawled a full Website or as many pages as it wanted in one day, so there'd be a clean copy. Other times, it might follow up later—many days later. Different crawl philosophies were used. The 1999 crawls do not contain a lot of images because we did not have enough bandwidth for text plus images. There were months when there was no crawling at all while the crawler was being rewritten” (Kahle 2002). In some cases this only shows empty content frames and some links (figure 29).



Figure 29. *Alexa.com through the Wayback Machine, Alexa 1999*

The Utilitarian Dream

Initially dreams about Internet Archive were rather utilitarian. The commercial use of the archive are described in the opening sentence of the article: “Bold efforts to record the entire Internet are expected to lead to new services” (1996):

In the end, our goal is to help people answer hard questions. Not “what is my bank balance?”, or “where can I buy the cheapest shoes”, or “where is my friend Bill?” - these will be answered by smaller commercial services. Rather, answer the hard questions like: “Should I go back to graduate school?” or “How should I raise my children?” or “What book should I read next?”. Questions such as these can be informed by the experiences of others.

Can machines and digital libraries really help in answering such questions? In the long term, we believe yes, but perhaps in new ways which would have importance in education and day-to-day life (Kahle 1996).

The technologies and the services seen to grow out of the Web archive would lead towards a reliable information interchange system based on electrons rather than on paper. The Wayback Machine is used to access documents that have disappeared from the Web, but the envisaged system to answer 'hard questions' is performed by other services on the 'live' Web, such as Amazon's book recommendation system. At present the Wayback Machine provides site search by URL and does not use the data in the archive in such a way to start answering these questions. In 2002, Kahle once again mentions his dream to create a machine that helps answering hard questions:

We want to grow our collections, but grow them in ways that they are useful to traditional library users—researchers, scholars, and the underserved. On the Web we can put tools and technology on top of collections—a search engine to answer harder questions. We can bring tools and information together in new ways that weren't possible before the Web. We think we have an archive, we want to build a digital library. We need partnerships to do this. We have collections but not a lot of finding aids. The top-level best thing for the community is universal access to human knowledge. It is within our grasp. We need to coordinate our efforts and just do it (Kahle, 2002).

This idea might have emerged from Kahle's previous project, Alexa.com, the Web Information Company, closely related to the Internet Archive (figure 30). Alexa Internet was founded by Brewster Kahle and Bruce Gilliat in 1996 and archived the Web from the start (Alexa History 2008). The Internet Archive, previously also called Alexa Archive (figure 31), donated its collection via Alexa crawls.²¹ The 1997 Internet Archive' Webmaster page the current Alexa.com Webmaster page are practically identical, both referring to the 'ia_archiver' crawler for instance (Alexa Webmasters Help 2009). The archived Web collection donated to the Internet Archive is based on a toolbar serving as an identifier to archive all pages visited by users with the toolbar installed.²² In return the toolbar offers Web surfers guidance on where to go next, based on the traffic patterns of its user community. The Alexa toolbar also offers context for each site visited, such as registration information, the number of pages, the number of sites referred at, its update frequency. In view of the enormous amount of Web usage data the Alexa toolbar collects from its users, Kahle's answer 'to hard questions' must have been based on this collection of data.

21 The location of the Wayback Machine itself has shifted around among several URLs during its first few months. Both <http://Web.archive.com> and <http://archive.alex.com> worked in the past, but at this moment they all redirect to www.archive.org. The URL <http://archive.alex.com> can be found in the archive. From 1998 till 24 September 2001, however, the message "There is no browsable Web service on this machine" is returned. On 24 September 2001 there is a slow but working Wayback Machine on the URL. From 10 November 2001 the URL is redirected to the Internet Archive. After 31 March 2002 the URL is no longer saved in the Archive and is currently no longer available online.

22 This made some users wonder about the origin of URLs in the Archive with actual search queries, including people's name, address, and personally identifiable information can be found in the Archive since (Planet 2008).

ALEXA Internet™

ALEXA Internet has created technology to crawl the Internet.


We have donated one copy of our crawl to the [Internet Archive](#) in an effort to help preserve our digital heritage.

Click here for more information about [the crawl](#).

Thanks, and stay tuned for future announcements.

 info@alexa.com

Figure 30. The first Alexa.com archived page in the Internet Archive, Alexa 1997



INTERNET ARCHIVE
WaybackMachine

Enter Web Address: All [Adv. Search](#) [Compare Archive Pages](#)

Searched for <http://archive.alexa.com> **17 Results**

* denotes when site was updated.
Material typically becomes available here 6 months after collection. [See FAQ.](#)

Search Results for Jan 01, 1996 - May 14, 2008												
1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
0 pages	0 pages	1 pages	2 pages	0 pages	11 pages	3 pages	0 pages	0 pages	0 pages	0 pages	0 pages	0 pages
		Dec 12, 1998 *	Jan 25, 1999 Feb 08, 1999		Sep 24, 2001 * Nov 10, 2001 * Nov 11, 2001 Nov 15, 2001 Nov 16, 2001 Nov 17, 2001 Nov 18, 2001 Nov 20, 2001 Nov 23, 2001 Nov 27, 2001 Nov 30, 2001	Jan 23, 2002 * Mar 29, 2002 Mar 31, 2002						

[Home](#) | [Help](#)

[Internet Archive](#) | [Terms of Use](#) | [Privacy Policy](#)

Figure 31. Archive.alexa.com in the Internet Archive, Archive.alexa.com 2009

The Academic Archives

As a new object of collection, Web documents have their own academic use and medium specificities compared to the libraries' traditional paper collections. Recalling the archival notion of archives as 'impartial evidence,' the transience of content, with the well-known error message "404 Document not found" admittedly makes the Web without archives "too unreliable" for academic citation (visiting professor at MIT, Carl Malamud in Kahle 1996). Web archives provide a rich source of information to study Webs over time. From the very beginning the archivist's project is dominated by the evidential nature of information on the Web. Documents such as Websites, news articles or discussion forums generally do not last longer than one year (Ntoulas et al 2004). On the other hand, Web content is more easily available for research purposes than paper collections. Where "historians have scattered club newsletters and fliers, physical diaries and letters, from past epochs, the

World Wide Web offers a substantial collection that is easy to gather, store, and sift through when compared to its paper antecedents" (1996).

Kahle regards the Internet as a medium that will become "a serious publishing system," making these archives and similar ones important to serve documents that may no longer be 'in print' (1996). This view on the Internet made science-fiction author Bruce Sterling remark: "Why *should* the Internet become a 'serious publishing system?' Who will give way first? Will the Internet somehow become a scholarly archive, or will scholarly archives become troves of uncatalogable spam and gibberish?" (Sterling 1996). In Kahle's article two dreams about the Internet Archive as important academic archives compete: on the one hand he foresees archives that include and preserve all written academic literature, on the other, archives serve as study object for historians and scholars, but may indeed include spam and gibberish, possibly evidence for studying social or cultural trends on the Web.

Legal Issues and Robots.txt

Legal issues such as copyright, privacy and intellectual property have prevailed the archiving project from the very beginning. The Internet Archive's solution to these issues is technical and provides Webmasters the possibility to 'opt-out' of being archived. On the first archived site of the Internet Archive, under 'Webmasters,' it reads: "We will not archive anything you request to remain private. All you have to do is tell us. How? By using the Standard for Robot Exclusion (SRE)" (1997). This technical solution uses 'robots.txt' files that Webmasters can manage on their Websites:

After retrieving any HTML file, we check for the presence of the NOINDEX, NOARCHIVE, and NO FOLLOW tags in the "<HEAD>" element of the document. If we find a NOINDEX or NOARCHIVE tag, we throw away the copy. If there is a NOFOLLOW tag, the robot will not follow any links we found on that page. The main advantage to this approach is that users can control access to their own data, without needing their site administrators to update "robots.txt" (Webmasters 1997)

If the archives crawlers do not visit a site, it is possible to 'opt-in' (figure 32).

Get Archived!

Do you want your site saved for posterity? Just fill out the simple form below.

Host:

Port:

Figure 32. Crawl me! Internet Archive Crawler, Internet Archive 1997

The Internet Archive strives to be both a digital library and a Web archive. The challenges are faced with technical solutions. The Internet Archive is collected, ordered and maintained by robots. Like search engines, the Internet Archive consists of a number of robots and servers automatically archiving most Webpages. Kahle proudly announces that: “For the first time, we can build a library that reads itself. You don't have the time to read all the books in a library — but your computer does” (Kahle 1997). Otherwise complicated legal issues, too, are thus delegated to software robots reacting to the robot instruction files on the sites visited.

The Internet Archive in 2009

The broad-sweep focus of the early Internet Archivists resulted in an archive that is not entirely complete, as is exemplified by empty frames and redirects to older archived versions or the live version of a Web page. However, without the Internet Archive, much of Web history as well as digital cultural heritage were lost. With the explosion of the Web, we risk living in what Danny Hillis referred to as the “digital dark age” (cited in Brand, 1999). The relevance of this ‘why archive the Web’-question, is a widely supported Web archivists’ and librarians’ point. Cofounder and director of the European Archives, Julian Masanès states, “cultural artifacts of the past have always had an important role in the formation of consciousness and self-understanding of a society and the construction of its future” (Masanès 2006: 1), the Web being the medium where contemporary culture in a large sense finds a natural form of expression. Publications, debate, creation, work and social interaction are “aspects of society that are happening or are reflected on the Web”²³ (Masanès 2006:1). The service Kahle has mentioned in 1996 to answer hard questions such as “should I go back to graduate school?” did not come out of the Internet Archive. Rather, the focus of the Internet Archive has further developed towards preserving Web-based cultural heritage. The how-question is approached in a medium-specific manner. Web content is identified, collected, preserved and made accessible with methods and techniques that mirror and thrive on the medium’s objects, dynamics and structures. As Brewster Kahle said, for the first time in history we have a system that can archive itself.

The cyberspace approach of the Internet Archive ensured the preservation of Web content. The Internet Archive currently contains almost 2 petabytes (2,000,000,000,000,000 bytes) of data and is growing at a rate of 20 terabytes per month. This exceeds the amount of text contained in the world's largest libraries (Internet Archive FAQ 2009). The current challenge is to make the archive scale with the evolving medium. The cyber spatial ideas from the early period have been confronted with national thoughts about the Web. In “A Fair History of the Web? Examining Country Balance in the Internet Archive” Mike Thelwall and Liwen Vaughan critically examine the Archive’s goal to index the entire Web. Is there is a national bias in what they cover? There are indeed large national differences in the Internet Archive’s coverage of the Web. The average age of sites in a country (i.e. old sites have more versions archived), hyperlink structures and the distribution of users that have the Alexa Toolbar installed are judged to be responsible for this uneven coverage.

²³ For social dimensions of networks see Castells (1996), Levy (1997), Hine (2000).

With the national turn, cultural heritage is thought to be best preserved at a national level. Brewster Kahle has adopted this vision in the current locative period by stating that it is “inconceivable for countries not to record their digital heritage. A lot of history is born digital” (2002). Working together with national archives, the Internet Archive tries to prevent the Web and its born digital materials from disappearing into the past. The current strategy to adapt to the changing medium is therefore to collaborate with national partners. Emerging national Web archiving initiatives, their models, methods, and their collaborative spirit is discussed in the following.

6. Archiving the National Web

With its origin in the period and spirit of cyber spatial thinking, the Internet Archive has in a certain way set the agenda in Web archiving. To what extent do the challenges facing the early archivists such as legal, technical and cultural relevancy issues still dominate the current field of Web archiving? Have the challenges changed, or the approaches to tackle them? The Internet Archivists' dream about an archive of the entire Web has transformed into a more decentralized national, as well as networked collaborative, archiving effort. The number of initiatives archiving the Web has grown exponentially and most of those archives have a national focus. Now the focus is mainly on how to agree to standards, methods and policies. Archival scientists as well as others examine the concept of archives, as well as the process of its institutionalization. Which institutions archive the Web and where did the trend of archiving the Web with a national focus emerge?

Historically libraries, archives and museums assumed the responsibility to preserve cultural heritage and to make it accessible. In most cases, collections were put on the Web and institutions started to construct Web archives. Most of them are government funded, so their collection has a national scope. The importance of these national archives has been emphasized in a combined JISC/NPO report as: "a strategy for digital preservation is part and parcel of any national information policy, and it should be integral to any investment in digital libraries and information superhighways" (PADI 2009).

In 1996 the Swedish and Australian national libraries began archiving a national collection of Web material. Since then, several national libraries followed (Masanès et al 2006: 41).²⁴ Most libraries have constructed topic-centric or thematic collections, such as archiving Websites related to national political parties in election time, major national or global events such as 9/11 or climate change. Electoral collections are archived by the Library of Congress' Minerva project (Schneider et al 2003) and the National Library of France's Election Archive (Masanès 2005). National archives also archive government's and local authorities' Websites, including the National Archives of Australia (Heslop et al 2002), the United Kingdom (Brown 2006), Canada and the United States (Carlin 2004). An early example of typical national Web archives is Kulturarw3, started in 1997 by the Swedish National Library (Arvidson et al 1998). This national archiving project tries to collect the national Swedish domain .se and Swedish pages linked from it but located in generic domains such as .com or .org (Masanès 2006: 41).

Some scientists have deemed the ephemeral nature of the Web as inappropriate for long-lasting referral and scientific verification, which requires access to the same data (Masanès 2006: 43). University libraries have therefore started archiving projects for the sole purpose of referencing. These are often topic or theme driven collections, like the Digital Archive for Chinese Studies (DACHS) at Heidelberg University in Germany (Lecher 2006), and the Dutch Archipol collection of political sites by the University of Groningen (Voerman et al. 2002). These thematic research driven collections often use human experts, i.e. researchers, to provide relevant sites. Certain archives

²⁴ Europe: Finland, Denmark, Norway, Iceland, France, Czech Republic, Slovenia, Italy, and Greece. Asia: Japan, China, and Singapore. USA: Library of Congress. The Library of Alexandria, Egypt, is one of the few providing online access, mirroring the Internet Archive.

base the selection of Websites to be archived on generic domains such as .gov (Cruse et al; Carlin 2004) or .edu (Lyle 2004).

It is important to underline that national Web archives, too, must contend with legal issues, such as copyright and intellectual property law. Peter Lyman states that “although the Web is popularly regarded as a public domain resource, it is copyrighted; thus, archivists have no legal right to copy the Web” (2002). At this time copyright law is enforced by national authorities, so that efficient archiving is best undertaken at national level and might therefore differ in each country. Web archives, set up by a national library, national archive or similar institution, try to develop local strategies for global digital preservation issues. In “The Role of National Initiatives in Digital Preservation” (2003), Margaret Hedstrom notes that archiving undertaken on a national level aptly addresses local concerns, and provides manageable solutions to problems which can become unmanageable if tackled on a global scale. By dividing responsibilities, effective national programs offer the advantage of manageability, whilst collaboratively benefiting from shared funding and research. Since all national institutions have similar technical problems and solutions to identify, fetch, store and make accessible Web archives, umbrella institutions, or consortiums, are being set up to collaborate. The International Internet Preservation Consortium (IIPC) is one of the central institutions sponsoring collaboration between national Web archives. In 2003 eleven national libraries together with the Internet Archive founded the IIPC to develop tools and collaborate on Web archiving activities. Currently, the IIPC group has 39 members; one of its main activities is developing a standardized and open source toolset to establish and maintain Web archives. This toolset comprises the Web crawler Heritrix, the archive format manipulation tool BAT, the access tool WERA, and the search engine NutchWAX.²⁵

The network location software, Issue Crawler, has been put to measure the international context the national Web archives operate in; its maps show how strong the ties are between the Web archiving initiatives. According to the National Library of the Netherlands, or Koninklijke Bibliotheek (KB) in Dutch, the IIPC is the most important Web archiving institution (Van Wijk 2009). All its members, mainly national libraries, have a Web archiving program. What do the hyperlinking strategies of the Web archiving initiatives reveal about the nature of their collaboration? The method used is called ‘inter-actor analysis,’ which means that links between a defined set of actors are counted and weighted. The crawl’s starting points are the IIPC members’ list (Appendix B).²⁶

The Issue Crawler extracts and follows all internal and external links from the starting points. Internal links point to another page within the same Website. Since the list consists of certain national libraries’ home pages, which are relatively large, the crawl depth is set to five levels.²⁷ Research has shown that crawling up to five levels deep is enough to reach 90 percent of a Website’s content (Baeza-Yates and Castillo 2004). Subsequently, only the external links to the sites within the starting

²⁵ Web archiving chain toolset (IIPC 2009)

²⁶ The list includes the IIPC

²⁷ National libraries have relatively large sites with many pages. In the field of interaction design, a rule of thumb for Web sites is that users should be able to visit all pages of a Web site within three clicks. National libraries, however, often have a number of collections and projects under the same domain and are therefore relatively large. For example, the Web archiving project of KB can only be reached after five clicks from the home page. The Issue Crawler has a limit of crawling three levels deep, but was set to depth five for this particular crawl.

point list are plotted on a map (figure 33). The size of the nodes indicates the number of links received from the network. Understandably the IIPC is one of the central actors, as its members are the starting points for the crawl. The Library of Congress, however, takes the lead. 21 of the national initiatives, mostly libraries, link to the Library of Congress' Web archiving program. There are 30 different top-level domains on the map, indicating a highly diverse population. 26 of the domains are country domains, i.e. at least 26 countries represented in this international network of national Web archiving institutions.

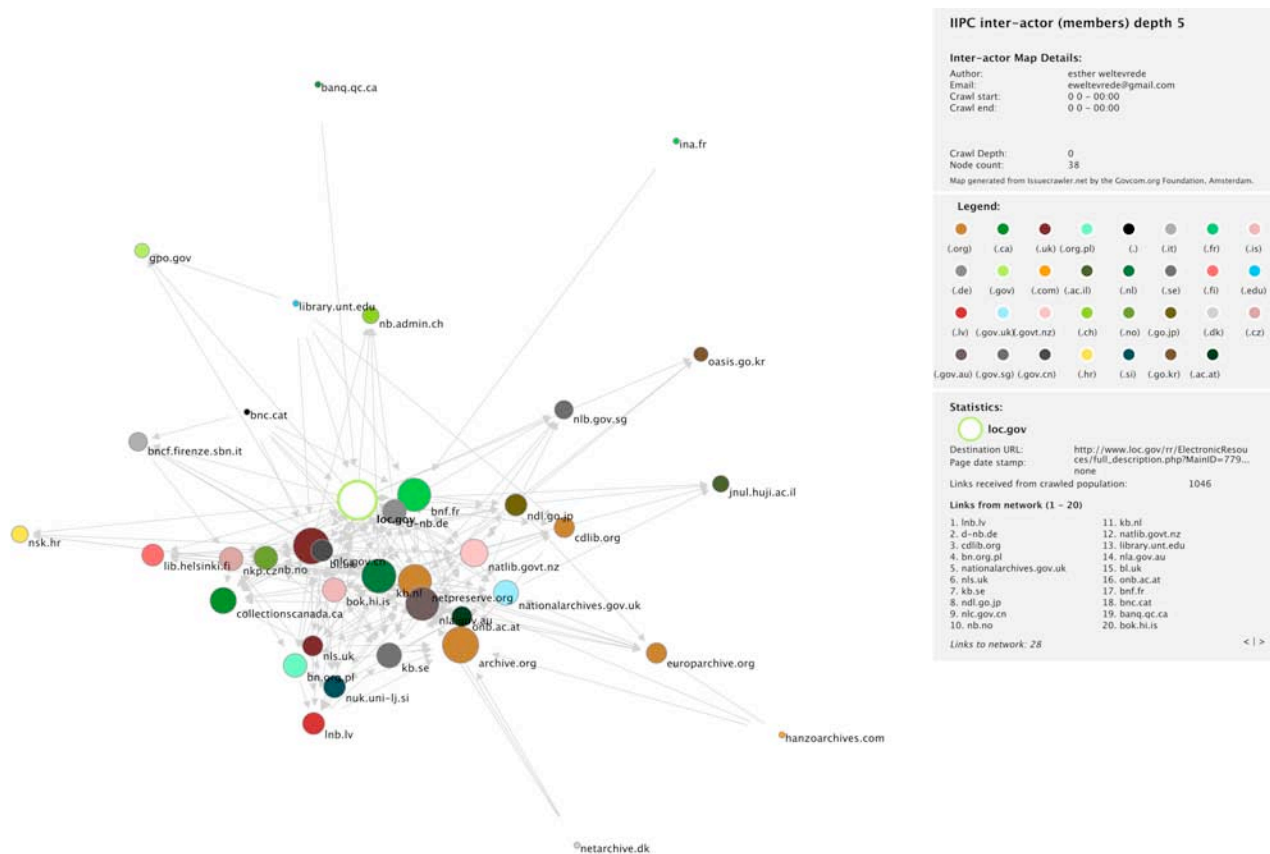


Figure 33. IIPC inter-actor (members)²⁸

The network is highly interlinked. In 2003 the United Nations Educational, Scientific and Cultural Organization (UNESCO) recognized the importance of saving digital cultural heritage in “UNESCO Charter on the Preservation of the Digital Heritage” (2003). One of the core goals of UNESCO is to promote international co-operation among its 193 member states and six associate members in the fields of education, science, culture and communication (UNESCO 2009). The UNESCO’s recognition with this Charter is important because it underwrites and supports both the Web archiving development to save Webs with a national focus, and the co-operation and development of common principles, policies, procedures and standards. The cooperation with relevant organizations and institutions is also encouraged in accordance with international norms and agreements (2003). In or-

²⁸ The original map is available online at Issuercrawler.net (IssueCrawler 2009)

der to gain a more specific perspective on who links to whom, some cases are highlighted in the following paragraphs.

The United States based Internet Archive is not nationally oriented, but promotes collaboration with several countries creating their own national archives to ensure the preservation of contents of historical relevance to their cultures. The Internet Archive is “affiliated with and receives support from various institutions, including the Library of Congress” (Internet Archive Legal 2009). According to the Issue Crawler map, however, the Internet Archive does only receive links from the network, but does not send any to the network, with the exception of one outgoing link to the Library of Congress (figure 34). According to the map, the Internet Archive is recognized as an important actor by the Web archiving institutions, but it does not affiliate with the others by linking back to other Web archiving projects and initiatives.

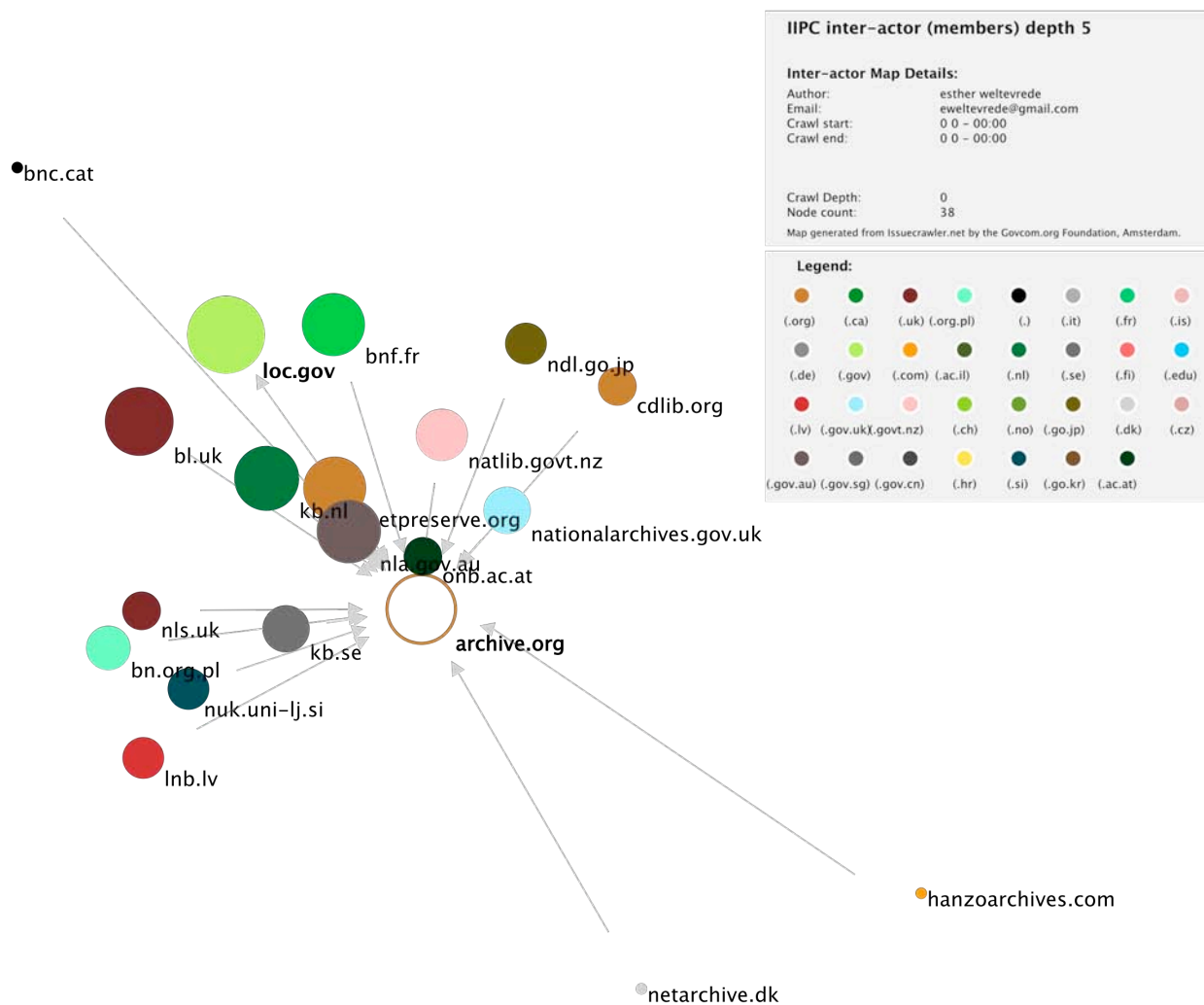


Figure 34. IIPC inter-actor (members) - Internet Archive

In a 2006 press release, the Amsterdam-based European Archive Foundation states that it strives to create a European digital library of cultural artifacts. “It acts as a technological partner for cultural

institutions to foster free online access to European cultural heritage and develops an open Web archive” (Masanès 2009). By partnering with the Internet Archive, the European Archive lays the foundation of “a global Web archive based in Europe” (Masanès 2009). Although the European Archive aims to be an umbrella institution for European cultural heritage, they are not yet recognized by the European national libraries or the Internet Archive (figure 35). Edwin van Huis, director of the Beeld en Geluid institute, says: “The European Archive is the only institute who tries to establish this public domain on a European level, not tied to nationalities, religion, cultural domains, governments, broadcasters or commercial partners” (Masanès 2009). So far, European Web archiving initiatives operate on a national and international level, not on a European level.

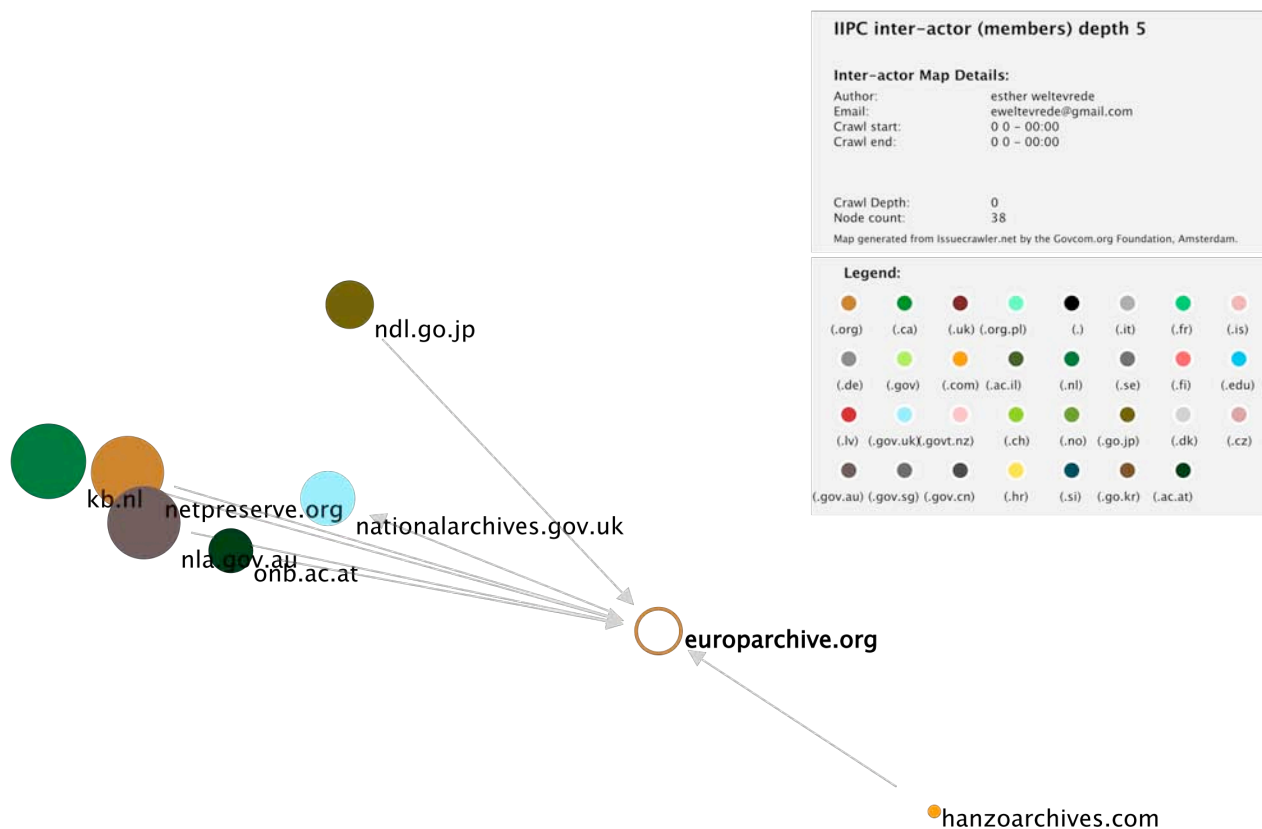


Figure 35. IIPC inter-actor (members) - European Archive Foundation

The KB is taken as an example for current Web archiving initiatives with a national focus. The KB is a national library, dedicated to save and make accessible a whole range of cultural heritage including books, magazines and newspapers as well as digital heritage including electronic journals and Websites. The KB is an active actor in the international collaboration between Web archiving institutions (figure 36). A large number of the network affiliates with the KB and vice versa.

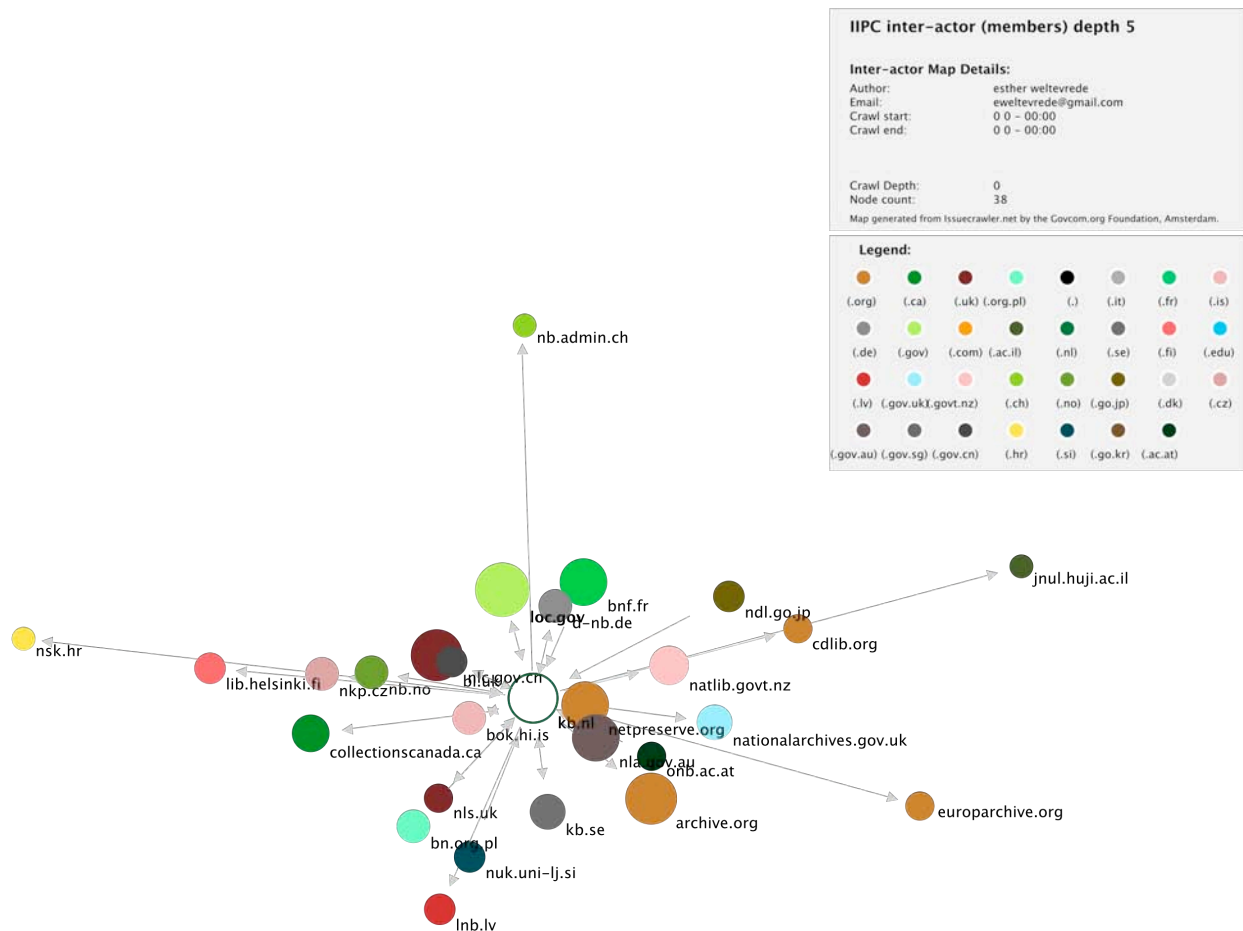


Figure 36. IIPC inter-actor (members) - Royal Library of the Netherlands

Web Archiving Models

In the following sections the various Web archiving models employed by national Web archives are described. Thereafter the KB is discussed further with these models in mind. The Web is not stratified along national lines, but needs to be configured nationally before it can be saved. How do the national Web archive initiatives deal with this complex construction of both nationalization as well as standardization of the Web archiving process in practical terms? How do current archivists decide what Web content is national and should be saved, and do they configure the Web along technical lines? The archivists' projects demonstrate national turn. The national Web archivists have to consider carefully which Web they aim to save for future generations and how they can delineate this Web. Here too, the technical methods used by the archivists as well as the technical apparatuses they create will be critically analyzed in a medium-specific manner. The aim is to strive to find out why and how the national Web archives look the way they do.

Since the late 1990s, many countries have been researching and experimenting with Web archiving models. In some cases the Web archives are supported by a legal deposit mandate for collecting national electronic resources, others collect publicly available sites belonging to a spe-

cific country's Web. A number of international partnerships explored and tested digital archiving theory and practice. Collaboration is necessary to prevent duplicates in archives across institutions. The Web archiving projects use various methodological approaches for discovery, acquisition and description of content. For a better understanding of the following models an understanding of the 'vertical' (or intensive) versus the 'horizontal' (or extensive) approach (Masanès 2005; see figures 37 and 38). Ideal archives are complete in a vertical fashion as well as in a horizontal fashion. However, "Web archiving is often a matter of choices, as perfect and complete archiving is unreachable" (Masanès 2005:77). Automated crawling can almost effortlessly discover and download millions of pages, but because of hyperlink structures it often implies that Websites are only partly preserved. Automated crawls have their limits and additional manual handling might be required to achieve vertical completeness. In topic or thematic collections such as DACHS and Archipol, the selection is made by a network of human experts to avoid overlooking pages that might not be linked to from the crawler's starting points. Web archiving projects therefore have to make a choice between an 'extensive' approach, if horizontal completeness is aimed at, or an 'intensive' approach, if vertical completeness is aimed at. This has direct implications on the quality and scalability of the archives.

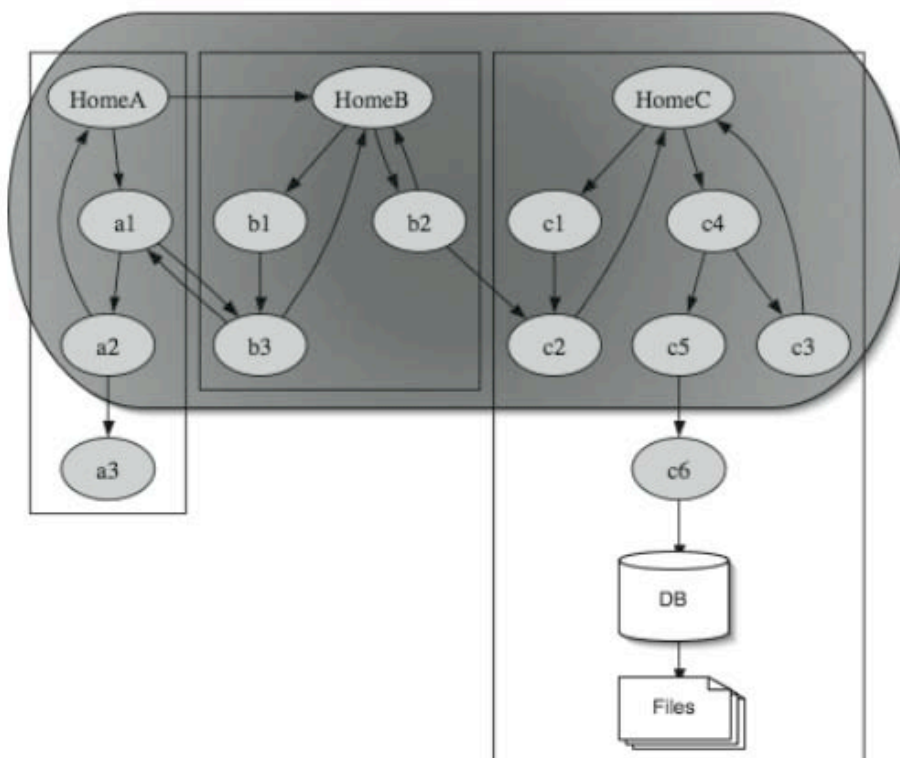


Figure 37. Extensive archiving (shaded area). Some pages are missing (a3, c6) as well as the 'hidden' part of sites (DB, Files), Masanès 2005

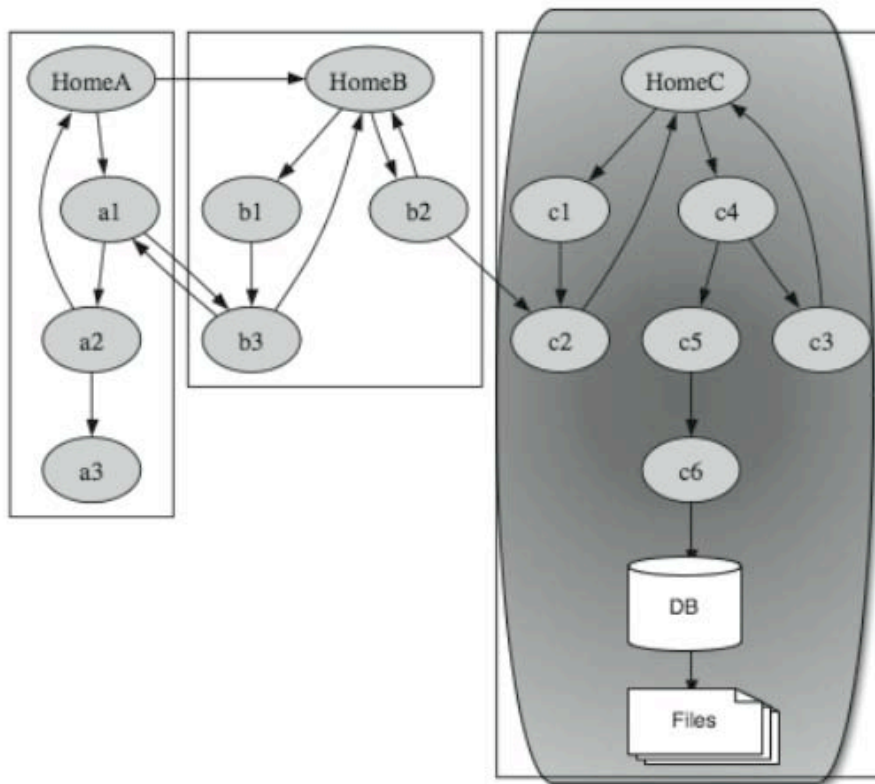


Figure 38. Intensive archiving (shaded area). Aims to collect fewer sites but collects deeper content, including potentially parts of the “hidden” Web, Masanès 2005

Several approaches or models have been developed for content collection, which can be categorized as follows:

Automatic Harvesting Model

Automatic harvesting is the extensive approach, whereby Web crawlers specially designed for this purpose download sites and online resources in broad sweeps of a national Web space. This is the approach of the Swedish Kulturarw3 (Cultural Heritage Cubed) and the Finish EVA project. This particular approach follows the technical apparatus of the country's assigned ccTLD and what those sites link to delineate the national Web. The Internet Archive's broad collection strategy is the most ambitious example, aiming at preserving the global Web's entire content (Day 2003; PADI 2008). Alexa's crawler uses a breadth-first approach, which means the crawler's search algorithm begins at the root node and explores all the neighboring nodes, and then adapts the crawl's depth according to the site's traffic (Burner 1997).

Selective Model

Selective or intensive approaches archive defined portions of a (national) Web space or specific types of resources. This is the case if crawlers use a site-first priority (Masanès, 2004) or if it is necessary to do a manual verification with additional archiving. This approach is more appropriate for 'deep Websites,' constructed dynamically from databases and content management systems, where crawlers do not have access to the full content (Masanès 2005). In some cases, repositories select

Web resources to be preserved and negotiate with site owners about inclusion, they then capture those resources with software for site mirroring or harvesting. The selection can be based on the resources' significance or quality. Rather than archiving full content, the selective approach aims at capturing a specific site or document by 'snapshots' at scheduled intervals. Examples of this approach are the National Library of Australia's Pandora archive, and the British Library (Day 2003; PADI 2008).

Thematic Model

A specific form of selective archiving that deserves its own category involves the archiving of Web content related to a particular theme or event. The Library of Congress's Minerva project used this approach when selecting a collection on the 2002 Elections and the Winter Olympics (PADI 2008).

Deposit Model

Publishers deposit Web-based material on a legal or voluntary basis into a repository. In Sweden the deposit of Web resources is legally required (Day 2003; PADI 2008). The deposit of Web resources by content owners is not well established in all countries, although there are several experimental projects. An example of a voluntary deposit scheme for electronic journals is by the KB in the Netherlands through agreements with publishers. The KB's deposit consists among others of Elsevier. The notion of 'national' is a rather arbitrary in this example as it is based on the place of publication, in the case of Elsevier the Netherlands.

Combined Models

The above-mentioned approaches are not mutually exclusive. A growing number of initiatives came to the conclusion that not one archiving model is entirely satisfactory for preserving national online heritage, among which the National Library of France, the National Library of Denmark and the National Library of New Zealand. These initiatives currently use a combination of selective, thematic and harvesting approaches for an optimal coverage of material. The selective approach is chosen to deal with technical complexity in Websites, as harvesting can be individually planned. This approach is especially suitable for the 'deep Web.'

Other Models

Each of the above-mentioned archiving models has advantages and disadvantage depending on the particular Web context applied to. New strategies, such as the 'by discipline' approach of some universities and research institutes do emerge, cf. the Digital Archives for Chinese Studies (DACHS) project of the University of Heidelberg. The Virtual Remote Control (VRC) program at Cornell University monitors changes to Websites over time (PADI 2008).²⁹

²⁹ A useful discussion on the relevant advantages and disadvantages of the various Web archiving strategies can be found in the study *Collecting and Preserving the World Wide Web : a Feasibility Study Undertaken for the JISC and Wellcome Trust*. Further introductions to Web archiving theory and practice are listed under Web archiving (PADI)

The outcomes of these models differ drastically. The automatic harvesting model privileges thinking in terms of the Web's technical apparatuses with their locative elements they use to configure a Web space according to national lines.³⁰ This model delegates most of the archiving process to automated gathering techniques and Web tools. The automatic harvesting approach is for instance used for domain-centric archiving.³¹ The selective approach is generally more time and expertise consuming (Day 2003; PADI 2008). It therefore tends towards the librarian approach, which is an editorial method to select content, but that entails that less Websites are saved. The thematic model can comprise both approaches depending on how material is selected in event or theme-driven collection: manually or e.g. by mapping hyperlink networks. The deposit-model is again a librarian approach, which generally entails a collection of (digitized) electronic journals.

KB's Definition of the Dutch Web

National Web archives have developed models for archiving the Web ranging from selective to automatic processing. The national library of the Netherlands (Koninklijke Bibliotheek – KB) is taken as a case study to critically examine outcomes of their approach, with the aim to find how the Dutch Web archive is saved for posterity. The KB's Web archive is not yet available online. A close reading of their technical documents, published papers and an interview are therefore the methods chosen for the analysis. The KB started its Web archiving project in 2006 by archiving one hundred Websites (KB - The Project 2009). With the experience of many initiatives in the field that have preceded the KB, the question arises, why does the KB save a limited selection of Websites of the Dutch Web for posterity?

According to Stichting Internet Domein Registratie Nederland (SIDN.nl), in 2008 the Dutch Web as defined in sites registered under the country domain .nl, consisted of over three million sites, the third largest country domain in the world (SIDN 2009). The KB did a survey to estimate the size of the Dutch Web. Not all registered names refer to a unique Website. For instance, <http://www.kb.nl>, <http://www.koninklijkebibliotheek.nl>, <http://www.koninklijke-bibliotheek.nl>, www.konbib.nl all refer to the same KB site. Based on surveys the Dutch Web is estimated to consist of 1,5 million unique Websites, covering at least 80 million Webpages. It should be noted, however, that these figures only refer to static sites and pages that can be indexed by search engines. Websites belonging to the deep Web that are not indexed by search engines are not included in these numbers. It is estimated that the deep Web exceeds the static pages by a factor 400 (Ras and Sierman 2006: 4).

30 For a discussion on how to demarcate a national Web space see Arvidson et al (2000), Abitbol et al (2002), Lamos et al (2002). For the studying of national Web characteristics see Baeza-Yates et al (2005 2x) and Gomes and Silva (2003). For the discovery and capture of content related to the same topic see Chakrabarti et al 1999; Bergmark 2002; Bergmark et al 2002; Qin et al 2004; Masanès ch 5 2006. Automated discovery and filtering is done using crawling techniques combined with page level appraisal of textual content, sometimes combined with link structure mining (Masanès 2006: 43).

31 Domain names are said not to follow rigid rules regarding names, functional specialization and organization, but rather to be formed by tradition (Liu and Albitz 1999; Koehler 1999). For archiving purposes it is important to note that all these Internet domain spaces are delegated by IANA and that each entity in charge can determine its policy with regard to the allocation and control of that space (Mueller 2002). This must therefore always be taken into account when making selections for archiving. Types of organizations are for instance not restricted in their choice of gTLD: .com is not necessarily commercial, .org not per definition non-profit. TLD entities in charge use an enormous difference in selection criteria. Moreover, archivists have to take into account that some entities change their TLD management and policy: .org and .net used to have restrictions before 1996, and .fr has reduced restrictions drastically since 2005.

The KB recognizes that the boundaries of a national Web can be defined in a number of ways: “the size depends on how the Dutch Web is defined exactly. Apart from all Websites hosted under the .nl domain, does it also include all .com, .net, .org, .eu domains that are written in Dutch or hosted in the Netherlands?” (Ras and Sierman 2006: 4). The KB decided to define the Dutch Web not in terms of country domain, but by a rather a broad concept of the Dutch Web domain. Their main criterion is: who is responsible for the content? Hence their criteria are not so much focused on “Dutch,” but rather on the “national aspect” of the content (Van Wijk 2009).³² The KB translated the ‘national aspect’ in four criteria (figure 39). In order to find out whether the KB approaches the medium technically, the question is: does KB’s definition of the national become before or after what is technically possible?

National aspect		
A	Website in Dutch, registered in the Netherlands	
B	Website in another language, registered in the Netherlands	
C	Website in Dutch, registered in another country	
D	Website in another language, registered in other country, topic aimed at the Netherlands	

Figure 39. Table of ‘national aspect’ criteria (see Appendix C for the extended selection criteria for Web archiving by the KB (in Dutch).

The four criteria for defining the national aspects contain three indicators: language, registration and the national aspect of a Website’s topic. Framing the Dutch Web in these terms means demarcating the Dutch Web along linguistic boundaries, the postal address of the site’s Webmaster and the nationality of topics or content on a site. For example, an international organization based in the Netherlands is included in the Dutch domain, as well as a Canadian site by or for Arnhem war veterans (Van Wijk 2009). The last indicator, mentioned in criterion D, is not a locative indicator. Also, determining the nationality of a topic automatically is thus far technically not possible. Strictly speaking, registration and language are also not technical indicators, but can be translated into technically feasible indicators.

Top-level domain .nl is a straightforward and effective indicator for the nationality of content on the Web, as for example the Swedish do, but is excluded from the list. It can therefore be concluded that the KB determines principles for the nationality of Web content before what is technically possible. This is illustrated by the non-technical topic criterion and the exclusion of technical indicators that can be used. The initial approach of the archivists at the KB is thus not a medium-specific one. However, the principles need to be translated into technically feasible ones for automated harvesting. A technical definition based on language and registration would however lead to a selection

³² Van Wijk 2009, personal correspondence with KB.

many times larger than the one hundred Websites the KB is archiving. The question remains, how did the KB end up with a limited selection of Websites to represent the Dutch Web in the archive?

The selective approach

Whereas most international initiatives began at an early stage and concentrated on website harvesting (an approach they are still following as a general rule), the KB has been emphatic in focusing its attention on the long-term storage of archived websites. The complexity of this task is the reason why the KB did not start web archiving until 2006. Since 2003, the development of an e-Depot system has provided the KB with an infrastructure that not only enables the electronic storage of articles from periodicals but also makes it possible to safeguard the archiving of websites (KB -The Project 2009).

The KB decided to follow a selective approach, i.e. to archive only a selection of the Dutch Web domain, which is divided into two phases. The first phase of the project lasted from January 2006 through June 2007 and aimed at acquiring knowledge and experience “into the various aspects of web archiving: the technique, the organization within the KB, the selection criteria, the costs, the legal consequences, storage, access and aspects having to do with digital preservation” (KB - The Project 2009). The second phase started in July 2007 and strived to set up a “web archiving infrastructure and embedding it within the existing KB workflow” (KB - The Project 2009). In late 2008 the operational and online-accessible Web archive was expected, but, in this time of writing, is still not online.

The KB states the choice for a selective approach has technical, legal, economic and institutional reasons (KB Selection 2009). With the current state of technology, there are too many Dutch Websites to ensure a quality harvest. The KB’s aim creating a high-quality source of research data is best achieved with a selection of Websites. The goal is to preserve entire Websites and not only the first three levels. In KB’s words: “Since the basic motive for web archiving is permanent storage, it does not seem wise to preserve only a limited portion of the websites. After all, we don’t store only the title pages of books” (KB-Selection 2009). A broad crawl or extensive bulk archiving would imply making snapshots of the Web, with strict limits on the number of files to be crawled per site and the amount of data stored. Even the entire Dutch domain, which is only a fraction of the total Web, is, according to the KB, too large to harvest in total and hence expensive to archive in its entirety (Van Wijk 2009). Moreover, the relatively limited selection contains sites offering a maximum of technical preservation challenges, thus creating the opportunity to learn without having to collect a vast amount of data (Ras and Sierman 2006: 5).

As national library of the Netherlands, the KB focuses on the long-term preservation of Dutch records. In countries with deposit legislation, national libraries are required by law to preserve national Websites. Unlike Denmark or France, the Netherlands does not have deposit legislation compelling site owners to provide copies to the KB. Instead, the KB must ask the site owner’s permission (Van Wijk 2009). Sites can therefore only be crawled after permission has been asked, thus making it difficult to crawl the complete Dutch Web. The Centre for Law in the Information Society (Centrum voor Recht in de Informatiemaatschappij; eLaw@Leiden) at Leiden University has done research for

the KB on how to best deal with intellectual property rights, such as the copyright, trademark right, neighboring rights, portrait rights and the database right. The recommendation is that site owners can opt-out instead of opt-in (Beunen en Schiphof 2006).

The selection principles are based on the KB's general collection policy, which aims to store and make accessible items about the Dutch language, culture and society. On top of this general KB collection policy, the Web archive also includes "output from government" and aim to preserve "Web 2.0 applications such as blogs and innovations and trends on the Web" (Appendix C). The primary selection stems from Websites with academic and cultural content, although innovative Websites as examples of current trends in the Dutch part of the Web are also considered (KB Selection 2009). The selection is based on the traditional cultural relevancy and, in the first phase, comprises just over a hundred carefully selected Dutch Websites.

The selection is made by experts, monitoring the Web within their field of expertise (Ras and Sierman 2006: 5). Experts at the department Academic Collection, which is part of the Expert Services and Collections Division, select Dutch Websites. As the name suggests, this is the KB's head division taking care of the field-specific aspects of library collections. The Websites are selected on the basis of the library's general collection plan. The second phase of the Web archiving project started in July 2007 was aimed at expanding the selection. This selection was initially made by DutchESS (Dutch Electronic Subject Service, 1993-2003) and results in about 400 Websites. Subsequently a deeper selection is made per subject field. In 2009 the fields of history, government administration and law are present in the Web archives. Taking into consideration the KB's collection policy, the exact sciences are underrepresented (Appendix C).

The core focus of the KB research program is the permanent preservation of digital documents,³³ which resulted in the e-Depot, the world's first long-term digital archiving system for academic publications (Ras and Sierman 2006: 8). After the sites are crawled and their quality is checked, they are stored in the e-Depot for preservation. The primary focus in this type of research is to make sure the archived Websites are still accessible in many years from now. On the Web, the presentation of content depends on browsers, but also on other aspects specific to a site like Flash or a specific kind of media player. Moreover, archiving essential software is required (Kiers 2007). Thus they decided to archive a limited and graspable set.

As demonstrated by the Issue Crawler map, the KB collaborates with international partners. The KB uses a set of open source tools specifically developed for Web archiving; they are brought together as an archiving tool set under the aegis of the IIPC. The open source tools include the Heritrix Web crawler, the NutchWax and Hadoop indexing tools, the Wayback Machine and WERA as search engines and Curator Tool for Web processing management. These tools enhance each other's functionality and together they are regarded as a complete archiving package (KB Technical Aspects 2009).

³³ The KB carries out research on two preservation strategies: migration and emulation. Migration focuses on the digital object itself and aims to change the object in such a way that software and hardware developments will not affect its availability. Emulation focuses on the hard- and software environment in which the object is rendered. The KB is developing both strategies as well as combinations of these methods, such as the Universal Virtual Computer for images (Ras and Sierman 2006: 8)

The KB uses Heritrix, the crawler developed by the Internet Archive (figure 40).³⁴ The Internet Archive use Heritrix for extensive crawling, while the KB asks Heritrix to crawl and collect a manually selected list of URLs. The Heritrix crawler wraps the target Websites' individual files in a kind of 'container,' to facilitate the management of the site's archived version. The individual files in this wrapping are described in terms of metadata, including information about the file format, the time and date of crawl, the file's size. The metadata is similar to those collected by the Internet Archive. Before the crawled sites are stored in the e-Depot they undergo a quality control, checking the harvested sites' completeness and quality and whether the links work. Furthermore, metadata is generated to describe the various file formats and their versions. The technical preservation metadata – file format, time and date of crawl, size of the file as well as versions – are important for future presentation (KB Projects 2009).

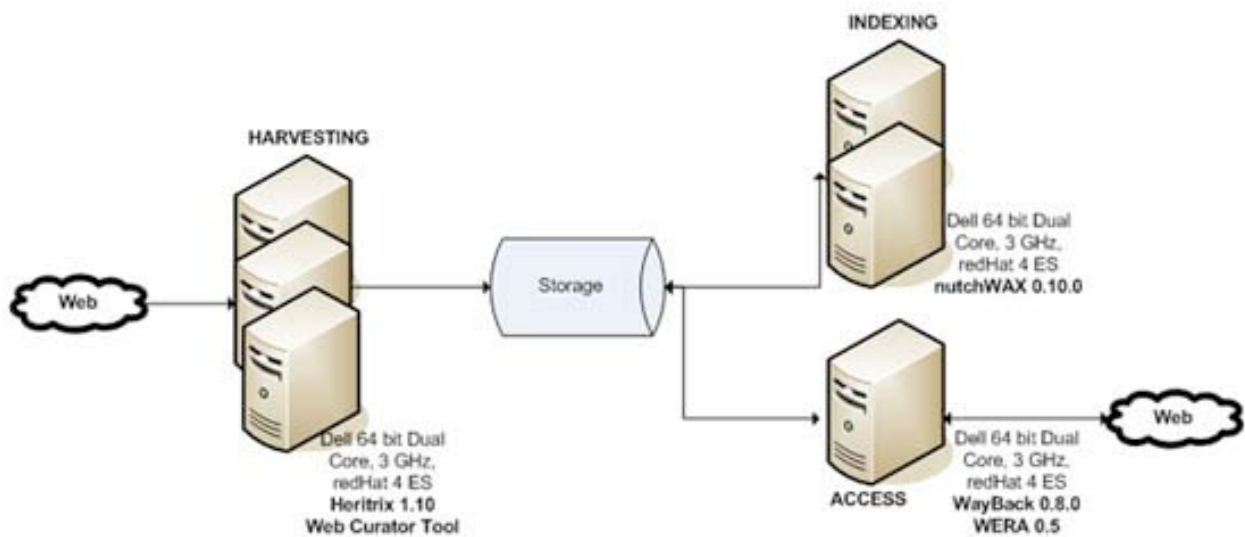


Figure 40. Tools in the KB archiving scheme, *KB Technical Aspects 2009*

Until recently the KB and the international IIPC partners paid little attention to the Web archives' interface. According to the KB there are two reasons why the interface received little attention. Firstly, most national initiatives started out by focusing on collecting and preserving Websites before thinking about how to make it accessible. Secondly, privacy and intellectual property legislation restrict public access to the Web archive (Van Wijk 2009). In 2007, the KB performed a user study for the Web archives. Users turned out to prefer a Google-like interface to the Web archive (Ras and Bussel 2007).

The KB is experimenting with various interface options for the Dutch Web archives. The first option is a Wayback Machine search by specific URL, like the Internet Archive. It means, however, that the searcher must know the URL to find anything. The second option is a full-text, Google-like search. A combination of NutchWax (for indexing and building search functionality) and the Wayback Machine (for presentation) is used to provide a full-text search. The IIPC group is collabora-

³⁴ Besides the standardization of Web archiving tools, the digital heritage field aims to develop and adhere to standards on other levels as well. Including metadata standards like Dublin Core and The Web ARChiving file format (WARC), which is the standard for collection, storage and retrieval of sites. Dublin Core: (ISO 2009); WARC (ISO 2009)

tively working on this search interface. This search interface, does have some disadvantages, however, in that, the free-text search often produces an enormous number of hits as there is no algorithm yet to determine ranking (Van Wijk 2009). Many of the hits are not very relevant to the original search inquiry. And the actual search engines, including Google,³⁵ cannot handle the time dimension well and this is one of the most important features for Web archives (KB Technical Aspects 2009). Currently the KB Web archive is not accessible online. The KB's goal is to make all Web archives full-text accessible to the end user (e.g. researcher). One of the possible search criteria is metadata such as the date of creation. The KB indexes the Web archives in the same way as other online accessible collections and does not use any ranking.

In terms of crawling, indexing, storing and searching Web archives, the KB's national Web archive and the Internet Archive adhere to equal or similar standards and they use similar tools. They differ in the way they approach the Web and hence in the specific technical arrangements they create. Contrary to the Internet Archive, but also the Swedish Kulturarw3, the KB has chosen to take a carefully planned and selective approach to archiving the Dutch Web. The resulting collection comprises a small collection of well-archived Websites, in contrast to Internet Archive, which has a large collection of partial Websites.

7. The Order of Things in the Digital

Web archives as technical arrangements are shaped by the period and spirit of their creation, mirroring dominant thoughts as well as technical developments. However, the dominance of the institutional context from which they emerge should not be underestimated. The Internet Archive that emerged from Web company Alexa, and their makers had a medium-specific approach from the start. They developed crawlers and sent them out to save as much of the Web possible and adjust in the process. Documents from the early Internet Archive mainly discussed the cultural and social relevance of saving Web content and concerns related to the medium, dominant in that period. The approach to tackle these issues, however, was technical. Legal issues for example, are also discussed in technical terms (i.e. robots.txt). The KB's Web archive is part of the national library of the Netherlands and takes a carefully planned expert approach to the archiving process. The collection policy and research focus are first shaped by the general library policy and research programs and subsequently translated into technical terms.

In this final chapter an effort is made to contribute to collection techniques for Web archiving are provided from a medium-specific approach, following archival principles. The approaches proposed adhere to the early archival theorists' principles and contribute to librarian categorization practices in the digital environment. How can the medium be repurposed to delegate parts of the collection process? How can the medium be made to work for the archivists?

Archivists save, order and make accessible information. Stemming from a tradition of preserving, cataloguing and indexing physical artifacts, Web archivists and librarians now face the challenge of preserving born digital material. As Niels Brügger notes in the line of Derrida, the process of archiving forms the object and, following a Foucauldian line of thinking, the resulting

³⁵ Google is working on a timeline, but this is still in labs (Google Timeline 2009)

shape informs what can be known with the archives. The Web and how it is considered influence what is saved and how it is made accessible to future generations, scientists and historians. The ordering dynamics of the Web entail that content may become meaningful in different ways. An important thing that is often overlooked in current Web archiving is the context in which it is embedded. As early as 1898, the Dutch archival trio advocated respect for the provenance, or 'birthplace,' of the records and the arrangement of the original record-keeping systems, in their times: the administrative context of state institutions (1898).

In the digital age the 'original order' of material has changed dramatically. Experts always had the task of ordering and structuring material. Physical things had, by their nature, to be assigned to one place. Knowledge material was ordered in a similar way, starting with Aristotle who categorized knowledge of beings in a hierarchical tree (figure 41). In this hierarchical model it is crucial that all things have a specific place in the hierarchy and have one single definition, validating that knowledge object' position in the whole order.



Figure 41. *Great Chain of Being*

Traditional libraries are exemplary of the way we order the things we know. Knowledge objects in libraries are materialized in books. A physical book can only have one specific place on the shelf. Librarians ordered the books in a library with metadata in indexes and catalogues so as to make books retrievable in more than just one way. In some cases this metadata structure imposed a strict categorization on knowledge. Exemplary is the Dewey Decimal System, which organized all knowledge in categories of ten (figure 42). Knowledge had to be squeezed into the perfect category.

000 General Works 010 Bibliography 020 Library Sciences 030 Encyclopedias, Almanacs 040 General Essays, Lectures 050 Periodicals 060 General Organizations 070 News, Journalism 080 General Collections 090 Manuscripts, Rare Books	100 Philosophy 110 Metaphysics 120 Epistemology, Causation and Humankind 130 Paranormal Phenomena 140 Philosophical School 150 Psychology 160 Logic 170 Ethics 180 Philosophy 190 Modern Western Philosophy	200 Religion 210 Natural Theology 220 Bible 230 Christian Theology 240 Christian Moral Theology 250 Local Religious Orders 260 Christian Social Theology 270 Church History 280 Christian Denominations 290 Non-Christian Religions	300 Social Sciences 310 General Statistics 320 Political Science 330 Economics 340 Law 350 Public Administration 360 Social Problems, Services 370 Education 380 Commerce, Communications 390 Customs, Etiquette, Folklore	400 Language 410 Linguistics 420 English and Old English 430 German 440 Romance, French 450 Italian and Romanian 460 Spanish and Portuguese 470 Latin 480 Hellenic, Classical Greek 490 Other Languages
500 Natural Sciences 510 Mathematics 520 Astronomy 530 Physics 540 Chemistry 550 Earth Sciences 560 Paleontology, Paleozoology 570 Life Sciences 580 Botanical Sciences 590 Zoological Sciences	600 Applied Sciences 610 Medical Sciences 620 Engineering 630 Agriculture 640 Home Econ., Cooking 650 Management 660 Chemical Engineering 670 Manufacturing 680 Specific Manufacturing 690 Construction Technology	700 The Arts 710 Civic and Landscape Arts 720 Architecture 730 Sculpture, Plastic Arts 740 Drawing, Decorative Arts 750 Painting and Drawings 760 Graphic Arts and Prints 770 Photography, Photographs 780 Music 790 Sports and Performing Arts	800 Literature 810 American Literature 820 English/Old English Literature 830 German Literature 840 French Literature 850 Italian/Romanian Literature 860 Spanish/Portuguese Literature 870 Latin Literature 880 Greek Literature 890 Other Literature	900 Geography/History 910 Geography and Travel 920 Biography and Genealogy 930 History of the Ancient World 940 European History 950 History of Asia, Orient and Far East 960 African History 970 North American History 980 South American History 990 Other History

Figure 42. Dewey Decimal Classification

In *Everything Is Miscellaneous: The Power of the New Digital Disorder* (2007) David Weinberger argues that knowledge no longer needs a definition, nor needs hierarchical categories to make sense. Rather, disorder is a feature of the digital environment. Following Jorge Luis Borges, it is an infinite library rather than a hierarchical classification of things. The order of things should be dependent on, and change according to, the context of the content object. Content objects can mean different things to different people in different contexts. In the digital, things can have infinite amounts of categories assigned to them and can be ordered in various layers atop of each other. The heterogeneity of categories and definitions assigned by different people in different contexts has taken over from the order of things created by experts. Two studies with a medium-specific approach, demonstrate how the role of the expert has changed in the digital environment.

The first focuses on the change in classification methods on the Web. The once popular Web directory in which experts categorized and ordered a carefully selected set of Websites became obsolete with the enormous growth of the Web, its popularity decreased because of the development of much more efficient search algorithms. In “The Googlization Question, and the Inculpable Engine,” Richard Rogers argues, “the burying of the directory in both Yahoo and Google signals a much larger transformation -- the demise of the expert human editors of the Web,” accompanied by “the rise of the back-end algorithm” (2009). “The demise of the directory” demonstrated that the editorial expert list has slowly vanished and is replaced by the algorithm. In, May 2008 Google’s directory was only found by querying Google (figure 44).³⁶

This study does not only demonstrate the changing role of the expert in the Web environment, but also witnesses a radical change in categorization schemes. The medium privileges methods that fit the medium’s ontology. This study demonstrates the changing role of the expert human

³⁶ Since the end of 2008 the directory has reappeared behind the ‘even more’ tabs. However, in the renewed version of the Google directory, the expert list is enhanced by the algorithmic Google PageRank to determine the position of sources: “Web Pages Ordered by PageRank. Unlike other directories that can only list Web pages alphabetically regardless of how good they are, the Web pages in the Google directory are ordered according to Google’s view of their importance. This means that the most relevant and highly-regarded sites on any topic are listed first ... not buried deep within a list of other pages.” (Google Directory Help 2009).

editor's methods. The manually selected and categorized list as well as the pre-determined fixed categories to categorize Web content make way for medium-specific methods, i.e. categorization schemes that continuously re-emerge in real-time from within the medium itself, ordered by natively digital constructs, such as PageRank using the hyperlink. The role of the expert has moved towards defining the rules for categorization built into natively digital constructs.

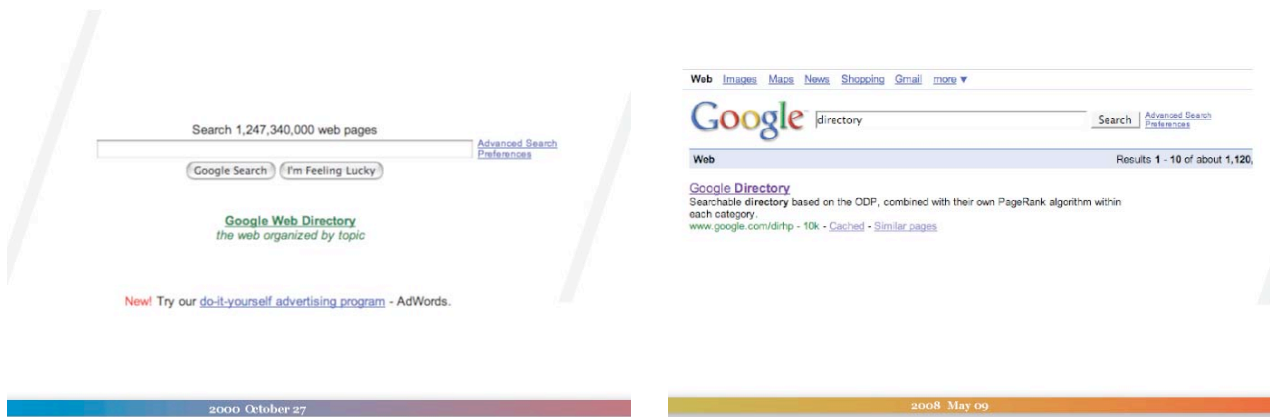


Figure 44. Selection Demise of the Directory, Digital Methods Initiative 2008³⁷

The second study revisits the 'original order' in Wikipedia. Comparing Encyclopedia Britannica and Wikipedia, Encyclopedia Britannica is a finite selection of knowledge, categorized and selected by experts, made accessible in a limited set of physical volumes, whereas Wikipedia on the other hand is an infinite and never finished knowledge-ordering project. Wikipedia does not present what we know as finished or perfect. Rather, the disordered and questionable nature of what is known is made apparent by notifications (figure 43). In "Wikipedia and the Vigilance of the Crowd" Web researcher Sabine Niederer makes an argument for the dependence of Wikipedians, in their knowledge production, on Wikipedia software robots (wikibots) (2009). The arrangement of expert Wikipedians and wikibots together can thus tweak and reorganize what is known. According to Niederer, the specific organizing power of the Wikipedia system is enabled by, what she calls the "technicity of content," including "tools and bots for editing, linking, combating vandalism, banning users, scraping and feeding content" (Niederer 2009).

Recalling Peter Scott, "the boundaries of the document have given way to a creative authoring event in which user and system participate. Only the context in which these virtual documents are created can give us an understanding of their content" (1990: 11). Niederer's contribution is, in Wikipedia, the creative authoring event is not only defined by expert users, but rather by users and bots. Wikibot Robbot, e.g. makes interwiki links to all different Wikipedia language versions of Wikipedia (2009). One single article can therefore be linked to and link back to multiple other articles in other language versions. And this is only one of the 889 wikibots contributing to the technicity of Wikipedia content (Wikipedia-Bots 2009). To gain an understanding of the content on Wikipedia, attention needs to be paid to its technicity of content.

37 For the full figure, as well as The Demise of the Directory, the movie, see the DMI wiki, (Rogers e.a., 2008)

Knowledge

From Wikipedia, the free encyclopedia



Figure 43. Notification in Wikipedia.org article on Knowledge, Wikipedia 2008

Either by bots or by algorithms, the human expert has to delegate work to the medium in order to cope with the enormous amount of information. Furthermore, the addition of the algorithm to the archiving and retrieval process allows a record to be in constant flux, but still ordered and retrievable, in the never finished knowledge-ordering project. Only the context in which Web material is created can give us an understanding of their content. How should the archiving and categorizing processes adapt to this new order in the Web environment? What is the role of the expert archivist on the Web?

In the Web archiving process archivists do not only have to take into account technical counterparts, but, according to the archival theory and practice tradition, also the social context counts in the archiving process. Web archivists and librarians have to work with knowledge objects that have a techno-socially constructed order, which is far from unambiguous. In the Web archives' interface, ideally both the librarian's categorization scheme and the archival original order of records meet. Approaching Web archives as knowledge ordering systems, they move away from principal ordering systems to arrangements taking into account a multiplicity of existing arrangements. Archival theorist Jerry Cook claims that this new order of things can be framed and approached with traditional archival concepts:

Original order should change from being viewed as the notion of a physical place for each record within a single series of records, to becoming instead a logical reflection of multiple authorship and multiple readership, where, for example, data may be united in multiple ways into new conceptual or virtual "orders" (or "series") for different transactions by different creators. A record will therefore belong to or reflect several series or original orders, not just one (Cook 1998: 48)

On the Web, it becomes relevant to view the arrangements that are in place, and how they organize and serve the Webs. This means moving away from the Web document and toward the authoring act or functional context surrounding the record. The challenge is not only to select the most relevant records, but also the prominence of the record in its original order, including relations between records, their functionality and place in a larger entity.

Medium-specific Collection Techniques

The revolutionary 'social turn' in archival principles replaced the state approach, which indexed official administrations. The question asked was, should not public opinion legitimize archival appraisal? When looking at the appraisal policy of the KB it is neither social, nor statist, but rather institutional. The question here asked, could not the medium be asked to legitimize archival appraisal?

In other words, how can the medium's computational power be used to select sources that are valuable to save for future generations?

Archival science has a rich tradition in focusing on the social dynamics related to records. Revisiting archival principles, the notion of archives is described as information generated as the 'by-product' of human activities (SAA 2002). On the Web records, too, are part of social dynamics. This study aimed to contribute to Web archival theory and practice by proposing techniques that adhere to the archival principles. On the Web, the social context of material is rather a techno-social context of material. In this section the argument is made that a natively digital object such as the hyperlink is the by-product of human activity: every link made is an act of association. The aim is to contribute to the collection techniques for Web archiving by proposing ways to find and map the techno-social context of 'digitally born' records by looking to their natively digital environment. The techniques proposed built on results from research with network location software, the Issue Crawler, which presents a collection method based on hyperlinks, and on previously discussed research project *The World According to Google*.³⁸

In "The Internet treats Censorship as Malfunction and Roots around it?" Richard Rogers argues that in a post-directory era, where the Google directory has been removed from the front page and Yahoo! is no longer the default search engine, relevance of documents follows from counting links and boosting sites either through freshness or through votes (2009: 236). As argued before, the expert librarian's role to make classification schemes of knowledge objects has partly been delegated to the algorithm. Classification schemes in the form of relevance measures are built into the algorithm. Every link made is like a vote (Google), or an act of association (Rogers 2009). Counting links therefore tells us something about what digitally born objects are considered important by a large number of people. The result pages of search engines constantly regenerate emerging classification schemes that contain both the social (many people casting votes) as well as the technological (the algorithm classifying sources by number of inlinks). Whether the algorithm shapes what many people consider relevant or vice versa, is arguable. What can be said, however, is that there are arrangements in place that go beyond the technical, or in other words, that are larger than the code of the algorithm. Rather, the ordering devices' algorithms reflect what sources many people find relevant. By privileging the medium's specific ways to recommend and give authority to documents on the Web, they go beyond capturing the Web's content objects by trying to capture the techno-social dynamics that make the Web into various national Webs.

A medium-specific approach to archival principles provides a way to start thinking of a collection of Websites situated in their techno-social context. In archival terms, digital methods provide means to delegate appraisal to the medium itself. The first technique proposes to use linking structures inherent to the Web. Traditionally crawlers use a 'snowballing' approach, the extensive method of crawling. This approach suffers from the fact that there is no way of knowing whether the sites found will be relevant. The Issue Crawler has implemented co-link analysis, which can be used to sample related URLs. Co-link analysis can be seen as a crossover between extensive and intensive archiving methods. It only keeps those sites in its result set deemed relevant by sites in the

³⁸ See chapter 3 *The Media of Location*.

network – through their linkages, a site only stays in the result set if at least two other sites link to it. This method thus finds related Websites by counting and weighting the number of links from a thematic-related set of starting points. When the starting points are for example the hundred sites the KB started their archiving project with, only those sites deemed relevant by a combination of those hundred sites are returned by the crawl, in addition to newly discovered sites deemed relevant by the network. The returned list of sites is ranked by the number of links they received from the crawled population, thus providing a network-specific ranking of the sources to be considered most relevant for the issue at hand. This ranking can then be used in the search interface of KB's archive.

Lastly, two techniques are proposed that build on The World According to Google research project and the tool used for that project: Generatenational.net. The first is a means to capture sites in a national Web, deemed the most relevant by the socio-technical context from which it emerges. Generate National does not only automatically count the number of results for TLD's in a country's Web, it also fetches the first 1000 results of the country's ccTLD or any gTLD in Google Region Search. The top 1000 of, for example .com, .org or .nl, sites are deemed most relevant according to votes casted by means of hyperlinks, can thus identified in one click. The results can subsequently be sent to the Issue Crawler to find out whether there is a hyperlink network within the top 1000 of a national Web. Recalling Ben-David's notion of the cyberstate and Halavais notion of the national, it counts and plots hyperlinks in order to see whether cyberstate borders (e.g. .nl), correspond with national communication flows (e.g. hyperlink networks within the .nl space).

The second technique proposed is an addition to Generate National to customize the tool for issue or event-driven collections within a national Web space. It builds on the event-driven collections by the Library of Congress in collaboration with the Internet Archive, including the presidential election collections, the 9/11 and Iraq war collection (Library of Congress-Web Archives, 2009). The addition to the tool comprises a search field to query a national Web for the defined issue or event. The resulting collection captures the Websites that are considered most relevant for an issue or event at a certain time.

The techniques proposed provide ways to delegate archival appraisal to the medium itself by capturing the socio-technical that is inscribed in the natively digital hyperlink. The technique using the Issue Crawler archives the by-product of human activity directly by working with the natively digital hyperlink. Generate National is built atop of technical arrangement Google. The resulting collection thus indirectly includes the by-product of human activity by building on Google's algorithm.³⁹ Building methods and tools atop of the natively digital and on Web arrangements is a means to capture Web culture and its socio-technical dynamics. Scheduling the above mentioned collection procedures in regular intervals is a means to capture the evolving Web dynamics of a national Web space over time in the archives. It provides ways to include both the Website as well as its prominence in a larger entirety in the archives.

³⁹ This particular method is built atop of Google. It is however recommended to and include and create methods for other Web arrangements in the collection process, such as Yahoo!, Hyves, and twitter.

Conclusion

The stakes are high for the humanities and social sciences. Archives traditionally serve as sites of knowledge production for various disciplines. There are however gigabytes of cultural heritage disappearing into the past every moment. The Web as a vast fluctuating amount of data is not an object simply there to be archived. This study started out by discussing cyberspace as an approach to think of Web space as well as Web space as ordered along national or linguistic lines. However, most of the early cyberspace is lost. Snippets of this early period can be found in the Internet Archive that started archiving in 1996. Web archives today face the important task to save more Web content from disappearing into the past. There are two larger points in this study. The first is a new way to think of Web space. The second strived to find out how and why current Web archives look as they do. The two points build up toward a contribution to Web archiving by proposing collection techniques from a medium-specific approach.

This study started out by discussing authors that observed the national turn on the Web and theorized this with their own approaches. Thinking in terms of language, users, flows or access, the counterintuitive trend observed by these authors is that content or users are clustered by nationality or language. It goes against intuition, indeed, because the Internet, with cyberspace as its conceptual framework, stood for universality and globalization instead of nationalization. The medium-specific approach to the national Web introduced, shows that the Web can be viewed as media of location, and more specifically, how the medium's native structures, objects and dynamics can be viewed as organized along national lines. Technical arrangements are defined as the systems that order Web content by technically defined measures.

The second larger point in this study strived to find out why how the archives look as they do. The initial hypothesis was that Web archives as technical arrangements are shaped by the period and spirit of their creation, mirroring dominant thoughts as well as technical developments. However, when looking for the technical arrangements of the archives, what I found was the dominance of the institutional context from which they emerge. This study thus identified two approaches that shape the archiving process. Firstly, the period and spirit from which the archiving projects emerge to shape the scope of the object of collection. When thinking about the Internet as cyberspace, a space larger than the Web becomes a targeted object of collection. Projects that emerged during the national turn, instead, put their archiving focus to a demarcated space of the Web, a national Web. Secondly, the institutional contexts from which the Web archives emerge drastically shape the archives by informing the approach to the process of archivization. The Web archiving project of the KB adhered to institutional principles and methods first; The Internet Archive, which emerged from a Web company, takes a technical approach.

The approach of the Internet Archivists' is medium-specific, but dates from cyber spatial ideas of the Internet. The order of things in the digital has changed. In the current state of the Web it has become relevant to view the arrangements that are in place, and how they organize and serve the Webs. This is a move away from archiving URLs and toward archiving the authoring act and functional context surrounding the digitally born material. It is a renewed focus on the interrelations, context and functionality of the born digital records, its creators, and its creation processes, wherever they occur. The current national turn provides the opportunity to revisit these medium-specific

methods to match the current state of the medium. Moreover, they must be revisited to save important digital born history from disappearing into the past.

Building on the two larger points of this study, a contribution was made to the field of Web archiving. The proposed techniques are ways to start thinking about what the Web archives could look like when archival principles meet with the national turn. Collection techniques that delegate appraisal and parts of the collection process to the medium itself were proposed. When thinking about national Webs, one can learn from how other technical arrangements that are native to the Web work with technological apparatuses that enable and constrain them to think along national lines within the Web.

The Web archivists' think of the Web as ephemeral transient medium; digital methods provide means to capture and save Web dynamics. Although the Web is also described as an archival or database medium, most of what happened on the Web in the past has vanished. Therefore, initiatives such as the Internet Archive and the KB are important, because they try to save part of the collective memory for future generations. But the kind of Web saved influences what can be studied with the Web archive and in what way. Digital methods provide new ways of saving the prominence of specific Websites or the dynamics around a certain issue over time and can serve as evidence of Web dynamics that will otherwise be lost.

To finalize, the implications of the recognition of the natively digital as object of study for archival sciences as well as Web research will be discussed, and I formulate suggestions for my future research project within the Digital Methods Initiative. The implications for Web archives as well as Web research are manifold. In digital preservation, the ontological distinction between 'digitized' and 'digitally born' material is made in the context of the Web environment. The digitally born refers to records that came into existence, and only exist, in the digital, while digitized refers to those objects that were migrated to the Web. The Digital Methods Initiative makes a further distinction between the 'digitally born' and the 'natively digital.' Although both 'born' in the digital environment, they belong to a different class of digital objects. The digitally born refers to file types such as .html or .jpeg, while the natively digital include locative and relational indicators such as the hyperlink, the IP-address and the tag. The digital born are 'content' containers, while the natively digital is what makes content 'networked' or what makes dynamic relations possible on the Web.

In the digital preservation context this distinction is important, because the digitally born as object of study considers the medium as transient, with records that are already lost. Taking the natively digital as object of study is a means to tame the medium. In other words, it is a means to embrace the medium's dynamics, its algorithms, its computational effects by making 'snapshots' of Web dynamics with custom made software tools. Digital methods exist by the grace of the ephemeral nature of the Web.

The natively digital as object of study means that this type of research is 'Web research' instead of 'Internet research.' Whereas Galloway's *Protocol* focuses on the low-level infrastructural layer of the Internet, which is a relatively timeless but also a static view of the Internet, the DMI focuses on the Web's dynamics, which, in terms of Internet research, might be located on the 'application layer' (figure 8). This type of medium-specific Web research thrives on the tension between the appreciation of the medium for its own specific merits, as well as everything that came before, it

builds on and shapes. In a similar vein as Matthew Fuller's Foucauldian approach to software - in terms of 'discursive formations' - in this study the technological, legal, economic and social influences shaping the medium par with the medium-specific methods shaped by the natively digital. Although the legal, economic and social aspects might be considered as external to the medium, this study chose to consider them to be embedded in the medium. Intellectual property legislation, for instance, became inscribed in the medium and shapes the order and place of Web content.

The study of technical apparatuses is a statist effort, while studying the results of Web arrangements is a national one. This first part of this study has focused on the technical methods used by technical arrangements to define the nationality of Web material and users. As demonstrated in chapter 3, the Webs as media of location, Web arrangements use technical apparatuses of the Internet infrastructure to order Web spaces along national lines. The focus in this study was on how technical arrangements re-territorialize cyberspace. In other words, because the objects of study were the technical indicators used by the technical arrangements, the effort was 'statist,' in terms of re-drawing borders.

The next step in this research project is moving towards national Web analysis, as was proposed in the collection methods for Web archiving. By studying the results of Web arrangements that order Web material and users nationally, claims can be made about the national. Instead of criticizing Web arrangement such as Google's information regime, the strategy proposed is to devise methods aimed at capturing these regimes for posterity. So that future Web research might look back at dominant information regimes of our times from their perspective.

References

- Anderson, B.R.O.G. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.
- Arvidson, A. et.al. (1998). "The Kulturarw Project-the Swedish Royal Web Archive." *Emerald Insight*.
- Baeza-Yates, R., and C. Castillo (2004). "Crawling the Infinite Web: Five Levels Are Enough." *Lecture Notes in Computer Science*.
- Barlow, J. P. (1996). "A Declaration of the Independence of Cyberspace." *Humanist-Buffalo* 56.
- Bearman, David (1990). "Multisensory Data and Its Management," in Cynthia Durance, ed., *Management of Recorded Information: Converging Disciplines*
- Ben-David, Anat (2008). "The Promised Cyberland: Does the State of Palestine Already Exist on the Web?" *Paper Presented at the Govcom.org Jubilee Summer Talks*.
- Bolter, J. D., and R. Grusin (1999). *Remediation: Understanding New Media*. MIT Press.
- Booms, H. (1987) "Society and the Formation of a Documentary Heritage: Issues in the Appraisal of Archival Sources." *Archivaria* 24.3: 69-107.
- Bradley, H. (1999). "The Seductions of the Archive: Voices Lost and Found." *History of the Human Sciences* 12.2.
- Brand, S. (1999). *The Clock of the Long Now: Time and Responsibility*. Basic Books.
- Brügger, N. (2005). "Archiving Websites." *General Considerations and Strate*.
- Burner, M. (1997). "Crawling Towards Eternity: Building an Archive of the World Wide Web." *Web Techniques Magazine* 2.5.
- Carlin, J. W. (2004). "Harvest of Agency Public Web Sites." *NARA Bulletin*.
<http://www.archives.gov/records_management/policy_and_guidance/bulletin_2005_02.html>
- Castells, M. (1999). "The Information Age: Economy, Society and Culture Volumes I, II, and III." *Journal of Planning Education and Research* 19.
- Chun, W. H. K. (2006). *Control and Freedom: Power and Paranoia in the Age of Fiber Optics*. MIT Press.

- . (2008). "The Enduring Ephemeral, or the Future Is a Memory." *Critical Inquiry* 35.1: 148-171.
- Cook, T. (1984). "From Information to Knowledge: An Intellectual Paradigm for Archives." *Archivaria* 19: 28-49.
- . (1998). "What is Past is Prologue: A History of Archival Ideas Since 1898, and the Future Paradigm Shift." *Archivaria* 43 <<http://www.mybestdocs.com/cookt-pastprologue-ar43fml.htm>>
- Cruse, P., Eckman, C., & Kunze, J. (2003). "Web-based Government Information: Evaluating Solutions for Capture, Curation, and Preservation." *California Digital Library*.
- Darnton, R. (2003). "How Historians Play God." *European Review* 11.03: 267-280.
- Day, Michael (2003). "Preserving the Fabric of Our Lives: a Survey of Web Preservation Initiatives." *UKOLN, University of Bath*.
- Derrida, J. (1996). *Archive Fever: A Freudian Impression*. University Of Chicago Press.
- Duff, Wendy, and Kent Haworth (1993). "The Reclamation of Archival Description: The Canadian Experience." *Canadian Archival Studies and the Rediscovery of Provenance*.
- Foucault, M. (1972). *The Archeology of Knowledge*. Trans. A.M. Sheridan Smith, Tavostock
- Fuller, M. (2003). *Behind the Blip: Essays on the Culture of Software*. Autonomedia.
- . (2008). *Software Studies: A Lexicon*, Leonardo Books
- Gilliland-Swetland, Anne (2000). "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment." *Council on Library and Information Resources*.
- Galloway, A. R. (2004). *Protocol: How Control Exists After Decentralization*. MIT Press.
- Gibson, W. (1984). *Neuromancer*. Ace Books.
- Goldsmith, J. L., and T. Wu (2006). *Who Controls the Internet? Illusions of a Borderless World*. Oxford University Press.
- Gomes, D., S. Freitas, and M. J. Silva (2006). "Design and Selection Criteria for a National Web Archive." *Lecture Notes in Computer Science* 4172.
- Greetham, D. C. (1996). "Textual Forensics." *PMLA* 111.1: 32-51.

- Halavais, A. M. C. (1998). "Measuring National Borders on the World Wide Web." Master Thesis, *The Department of Communication at the University of Washington*
- Hedstrom, M. (28-30 September 1998). "The Role of National Initiatives in Digital Preservation." *Guidelines for Digital Imaging: Papers Given at the Joint National Preservation Office and Research Libraries Group Preservation Conference in Warwick* 85-7.
- Heslop H. et al. (2002). "An approach to the Preservation of Digital Records." *National Archives of Australia*.
- Hine, C. (2000). *Virtual ethnography*. Sage Publications Inc.
- Hölscher, Christoph, and Gerhard Strube (2000). "Web Search Behavior of Internet Experts and Newbies." *Institut für Informationssysteme und Computer Medien*. 8 Mar 2009
 <[http://www.iicm.edu:8000/thesis/cguetl_diss/literatur/Kapitel02/References/Hoelscher et al. 2000/81.html?timestamp=1194675006470](http://www.iicm.edu:8000/thesis/cguetl_diss/literatur/Kapitel02/References/Hoelscher_et_al._2000/81.html?timestamp=1194675006470)>.
- Jenkinson. Hilary (1922). *A Manual of Archive Administration*. P. Lund, Humphries and Co.
- Lecher, H. E. (2004). "Informant Networks, Alarm Systems, and Research Contributors. Selection and Ingest Process for the Digital Archive for Chinese Studies. *Paper presented at the Archiving Web Resources Conference, Issues for Cultural Heritage Institutes*.
- Lilley, Robert, et.al (2006). "White Paper GPS Backup For Position, Navigation and Timing Transition Strategy for Navigation and Surveillance." *Federal Aviation Administration ATO-W Navigation Services Cooperative Agreement 06-G-001*.
- Lovink, Geert (2009). "Internet, Globalization and the Politics of Language." Pre-publication.
- Lyle, J. A. (2004). "Sampling the Umich.edu domain." *Paper Presented at the 4th International Web Archiving Workshop*.
- Lyman, P. (2002). "Archiving the World Wide Web." *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. 38-51.
- Lynch, C. et al. (1995). *CNI White Paper on Networked Information Discovery and Retrieval*.
- Mackenzie, A. (1997). "The Mortality of the Virtual: Real-time, Archive and Dead-time in Information Networks." *Convergence* 3.2.

- Manovich, L. (2001). *The Language of New Media*. The MIT Press.
- Masanès, Julian (2004). "Site-first Priority: Implementing the Frontline." *International Internet Preservation Consortium*.
- . (2006). *Web Archiving*. Springer-Verlag, Inc. Secaucus.
- . (2005). "Web Archiving Methods and Approaches." *Library Trends* 54: 72-90.
- McGuinness, D. L. et al. (2006). "Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study." *Proceedings of the Workshop on Models of Trust for the Web*.
- Miller, D., and D. Slater (2000). *The Internet: An Ethnographic Approach*. Berg Publishers.
- Muller, Samuel, Johan Feith and Robert Fruin (1898). *Manual for the Arrangement and Description of Archives*. Society of American Archivists.
- Nesmith, T. (1993). "Archival Studies in English-speaking Canada and the North American Rediscovery of Provenance." *Canadian Archival Studies and the Rediscovery of Provenance*.
- Niederer, Sabine (2009). "Wikipedia and the Vigilance of the Crowd." *Paper Presented at the Digital Methods Initiative Progress Meeting*.
- Norton, M. C., and T. W. Mitchell (2003). "Norton on Archives: The Writings of Margaret Cross Norton on Archival & Records Management." *Society of American Archivists*.
- Ntoulas, A., J. Cho, and C. Olston (2004). "What's New on the Web?: The Evolution of the Web from a Search Engine Perspective." *Proceedings of the 13th International Conference on World Wide Web*. ACM New York: 1-12.
- Osborne, T. (1999). "The Ordinarity of the Archive." *History of the Human Sciences* 12.2
- Rheingold, H. (2007). "Using Participatory Media and Public Voice to Encourage Civic Engagement." *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*: 97-118.
- Rogers, Richard (2009a). "A New Media Approach to the Study of State Internet Censorship." *Pre-publication at Govcom.org*.
<http://govcom.org/publications/full_list/Rogers_in_Parikka_Spam_book_optimized.pdf>
- . (2004). *Information Politics On The Web*. MIT Press.

---. (2009b.). "The Googlization Question, and the Inculpable Engine." *Paper Presented at Digital Methods Initiative Progress Meeting*.

<http://govcom.org/publications/full_list/rogers_inculpable_engine_27Dec2008.pdf>

---. (2007). "The Politics of Web Space." *Prepublication at Govcom.org*.

<http://govcom.org/publications/full_list/rogers_politics_web_space_2008_pre.pdf>

Said, Edward (1979). *Orientalism*. Vintage Books.

Samuels, H. W. (1992). *Varsity Letters: Documenting Modern Colleges and Universities*. Scarecrow Press, Inc.

Schellenberg, T. R. (1956). "The Appraisal of Modern Public Records." *Bulletin of the National Archives* 8.

Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). "Building Thematic Web Collections: Challenges and Experiences from the September 11 Web Archive and the Election 2002 Web Archive." *Paper Presented at the 3rd ECDL Workshop on Web Archives*.

Scott, J. C., (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Am Anthropol Assoc.

Shirky, C. (2008). *Here Comes Everybody: How Digital Networks Transform Our Ability to Gather and Cooperate*. Penguin Press.

Stvilia, B. et al. (2005). "Information Quality Discussions in Wikipedia." *Proceedings of the 2005 International Conference on Knowledge Management*. 101-113.

Sunstein, C. R. (2006). *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press.

Tessler, Joelle (15 July 2000). "Online Auctions of Nazi Items Sparks Debate Issue: National Laws on the Global Web." *San Jose Mercury News*.

Thelwall, M., and L. Vaughan (2004). "A Fair History of the Web? Examining Country Balance in the Internet Archive." *Library and Information Science Research* 26.2: 162-176.

Tuters, M., and K. Varnelis (2006). "Beyond Locative Media: Giving Shape to the Internet of Things." *Leonardo* 39.4: 357-363.

Velody, I. (1998). "The Archive and the Human Sciences: Notes Towards a Theory of the Archive." *History of the Human Sciences* 11.4.

Voerman, G., Keyzer, A., Hollander, F. D., & Druiven, H. (2002). Archiving the Web: Political Party Web Sites in the Netherlands. *European Political Science*.

Wardrip-Fruin, N., and N. Montfort (2003). *The New Media Reader*. MIT Press.

Weinberger, D. (2007). *Everything Is Miscellaneous: The Power of the New Digital Disorder*. Times Books.

Wijk, Caroline van (12 January 2009). "7 vragen KB." Esther Weltevrede. Email interview.

Wilson, I. E. (1995). "Reflections on Archival Strategies." *American Archivist* 58: 414-429.

Withers, C. W. J. (2002). "Constructing the Geographical Archive." *Area* 34.3: 303-311.

Woolgar, S. (2002). *Virtual Society?: Technology, Cyberbole, Reality*. Oxford University Press.

Web References

Agarwal, Amit (24 Mar 2008). "YouTube Video Not Available in Your Country? How to Watch Blocked Videos." *Digital Inspiration*. 8 Mar 2009 <<http://www.labnol.org/Internet/video/youtube-blocked-video-not-available-in-your-country/2680/>>.

Alexa - "History (2009)." *Alexa*. 8 Mar 2009 <<http://www.alexa.com/site/company/history>>.

Alexa Top Sites (2009). *Alexa*. 8 Mar 2009 <http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none>.

Alexa - "Webmasters Help" (2009). *Alexa*. 8 Mar 2009 <http://www.alexa.com/site/help/Webmasters#crawl_site>.

Boyle, Alan (10 Aug 1997). "Archiving the Internet for Posterity." *MSNBC in the Internet Archive*. 8 Mar 2009 <<http://Web.archive.org/Web/19970810233147/www.msnbc.com/news/60659.asp>>.

Beunen, Annemarie en Tjeerd Schiphof (2006). "Legal Aspects of Web Archiving from a Dutch Perspective." *The National Library in The Hague*. <http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_Legal_Aspects_WebArchiving_EN.pdf>

KB - DTD, "Bibliographic DTD of the Koninklijke Bibliotheek (23 Mar 2001)." *Koninklijke Bibliotheek*. 8 Mar 2009 <http://www.kb.nl/hrd/catalogus/kb_dtd.txt>.

Borra, Erik, and Michael Stevenson (12 Sep 2008). "Repurposing the Wikiscanner." *Digital Methods Initiative*. 8 Mar 2009 <<http://wiki2.IssueCrawler.net/twiki/bin/view/Dmi/WikiScanner>>.

Crawler Manual - "6. Configuring Jobs and Profiles (2009)." *The Internet Archive Crawler Manual*. 8 Mar 2009 <http://crawler.archive.org/articles/user_manual/config.html> and "8. Analysis of Jobs (2009)." *The Internet Archive Crawler Manual*. 8 Mar 2009 <http://crawler.archive.org/articles/user_manual/analysis.html>.

Cunningham, Michael (27 Jan 1997). "Brewster's Millions." *The Irish Times in the Internet Archive*. 8 Mar 2009 <<http://Web.archive.org/Web/19990117002422/www.irish-times.com/irish-times/paper/1997/0127/cmpl.html>>.

Dodge, Martin, and Rob Kitchin (2008). *Atlas of Cyberspace*. 8 Mar 2009 <<http://www.kitchin.org/atlas/contents.html>>.

Find IP-address (2008). "IP-addresses - IP Range | IP Location - IP-address Ranges". *Find IP-address*. 8 Mar 2009 <<http://www.find-ip-address.org/ip-country/>>.

Fuller, M. (2006). "Software Studies Workshop." *Piet Zwart Institute*, 19 Jan 2009 <<http://pzwart.wdka.hro.nl/mdr/Seminars2/softstudworkshop>>.

Garb, Rachel (7 Jul 2008). "More Transparency in Customized Search Results." Official Google Blog. 8 Mar 2009 <<http://googleblog.blogspot.com/2008/07/more-transparency-in-customized-search.html>>.

Google - "Corporate Information - Company Overview" (2009). Google 2009. 8 Mar 2009 <<http://www.google.com/corporate/>>.

Google Timeline (2009). "Civil Rights Movement View:timeline". Google Timeline. 8 Mar 2009 <<http://www.google.com/views?q=civil+rights+movement%20view%3Atimeline&esrch=RefinementBarTopViewTabs>>.

Google - "Web Directory Help" (2009). Google. 8 Mar 2009 <<http://www.google.com/intl/en/dirhelp.html#pagerank>>.

Griffith, Virgil (2009). "WikiScanner." WikiScanner Virgil. 8 Mar 2009 <<http://wikiscanner.virgil.gr/>>.

Helmond, Anne (22 Sep 2008). "Review: Software Studies a Lexicon Edited by Matthew Fuller." Anne Helmond. 8 Mar 2009 <<http://www.annehelmond.nl/2008/09/22/review-software-studies-a-lexicon-edited-by-matthew-fuller/>>.

IBM - "News - Researchers Map the Web" (2009). *IBM Almaden*. 19 Jan 2009 <http://www.almaden.ibm.com/almaden/Webmap_release.html>.

ICPSR - "Digital Preservation Tutorial" (2007). *Cornell University Library*. 8 Mar 2009 <<http://www.icpsr.umich.edu/dpm/dpm-eng/terminology/repository.html>>.

IIPC - "Software - Downloads" (2008). *International Internet Preservation Consortium*. 8 Mar 2009 <<http://www.netpreserve.org/software/downloads.php>>.

Internet Archive - About (2009). "Initiatives, About Internet Archive". *Bibliotheca Alexandrina*. 8 Mar 2009 <<http://www.bibalex.org/english/initiatives/Internetarchive/about.htm>>.

Internet Archive - "Building a Library for the Future" (11 Oct 1997). *The Internet Archive in the Internet Archive*. 8 Mar 2009 <<http://Web.archive.org/Web/19971011064403/http://www.archive.org/index.html>>.

Internet Archive - "Details: Brewster Kahle speaks at the Library of Congress" (13 Dec 2004). *Internet Archive*. 8 Mar 2009 <http://www.archive.org/details/cspan_brewster_kahle>.

Internet Archive - "Forums: an Open Question to Archive/Alexa" (31 Mar 2008). *The Internet Archive*. 8 Mar 2009 <<http://www.archive.org/iathreads/post-view.php?id=185671>>.

Internet Archive - "Frequently Asked Questions" (2009). *The Internet Archive*. 8 Mar 2009 <<http://www.archive.org/about/faqs.php>>, <<http://www.archive.org/about/faqs.php#10>>, <<http://www.archive.org/about/faqs.php#3>>, <<http://www.archive.org/about/faqs.php#202>>.

Internet Archive - "Homepage" (2009). *The Internet Archive*. 8 Mar 2009 <<http://www.archive.org/index.php>>.

Internet Archive - "Legal: Affidavit" (2009). *The Internet Archive*. 8 Mar 2009 <<http://www.archive.org/legal/affidavit.php#forum>>.

Internet Archive - "Webmasters" (11 Dec 1997). *The Internet Archive in the Internet Archive*. 8 Mar 2009 <<http://Web.archive.org/Web/19971211124750/www.archive.org/Webmasters.html>>.

ISO - Dublin Core (2009). "ISO 15836:2003 - Information and Documentation - The Dublin Core Metadata Element Set". *International Organization for Standardization*. 8 Mar 2009
<http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37629>.

ISO - WARC (2009). "ISO/PRF 28500 - Information and Documentation -- WARC File Format". *International Organization for Standardization*. 8 Mar 2009
<http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717>.

Kahle, Brewster (4 Nov 1996). "Archiving the Internet." *The Internet Archive*.

---. (16 Jan 2007). "Analysis of Search Activities of Users to Identify Related Network Sites". *Google Patents*. 8 Mar 2009 <<http://www.google.com/patents?hl=en&lr=&vid=USPAT7165069&id=ebL-AAAAEBAJ&oi=fnd&dq=archive+b-kahle>>.

---. (2002) "Editors' Interview - The Internet Archive". *RLG DigiNews*. 8 Mar 2009
<<http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070513/viewer/file993.html>>.

---. (2004) "Universal Access to All Knowledge" *ITConversations*.
<<http://itc.conversationsnetwork.org/shows/detail400.html>>

KB - About "Over de KB" (2009). *Koninklijke Bibliotheek*. 8 Mar 2009
<<http://www.kb.nl/menu/kb.html>>.

KB - "Legal Aspects" (2009). *Koninklijke Bibliotheek*. 8 Mar 2009
<http://www.kb.nl/hrd/dd/dd_projecten/Webarchivering/juridisch-en.html>.

KB - "Selection" (2009). *Koninklijke Bibliotheek*. 8 Mar 2009
<http://www.kb.nl/hrd/dd/dd_projecten/Webarchivering/selectie-en.html>.

KB - "The Project." *Koninklijke Bibliotheek*. 8 March 2009
<http://www.kb.nl/hrd/dd/dd_projecten/Webarchivering/project-en.html>

Kiers, Bart (2007) "Web Archiving within the KB and Some Preliminary Results with JHove and DROID." *Koninklijke Bibliotheek*. 9 Mar 2009.
<www.kb.nl/hrd/dd/dd_projecten/Webarchivering/documenten/IIPC-PWG-Webarchiving-JHove-DROID-test.pdf>

Kirschenbaum, Matthew (2007). "A Printing House in Hell." *University of Maryland*. 8 Mar 2009
<http://www.otal.umd.edu/~mgk/blog/archives/2007_06.html>.

Library of Congress - "Technical Information" (6 Mar 2008). *Library of Congress Web Archives Minerva*. 8 Mar 2009 <<http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-techinfo.html>>.

Library of Congress - "Alexa Internet Donates Archive of the World Wide Web To Library of Congress (13 Oct 1998)." *The Library of Congress*. 8 Mar 2009 <<http://www.loc.gov/today/pr/1998/98-167.html>>.

Library of Congress - "Web Collections" (2009). *The Library of Congress*. 8 March 2009, <<http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html>>

Manovich, L. (2008). *Software Takes Command*. Softbook. <<http://www.softwarestudies.com/softbook>>

Marcum, D. B. (1998). "We Can't Save Everything." *New York Times* 6.

Martin, R. (2002). "Sharing the Wealth." *RLG Members' Forum, Washington, DC*.

Masanès, Julian (28 Sep 2006). "Official Launch of the European Archive Foundation." *European Archive*. 8 Mar 2009 <<http://www.europarchive.org/launch-official.php?PHPSESSID=658a51b06d19de25bee2442b7aee853f>>.

Merriam-Webster - "Archive" (2009). *Merriam-Webster Online Dictionary*. 8 Mar 2009 <<http://www.merriam-Webster.com/dictionary/archive>>.

OAIS - "Glossary" (2007). *Digital Preservation at ICPSR*. 8 Mar 2009 <<http://www.icpsr.umich.edu/cocoon/DP/glossary.xml?token=OAIS>>.

OpenNet Initiative (2009). "Regional Overviews". *OpenNet Initiative*. 8 Mar 2009 <<http://opennet.net/research/regions>>.

PADI - "National Strategies" (2009). *Preserving Access to Digital Information*. 8 Mar 2009 <<http://www.nla.gov.au/padi/topics/68.html#org>>.

PADI - "Web Archiving" (2009). *Preserving Access to Digital Information*. 8 Mar 2009 <<http://www.nla.gov.au/padi/topics/92.html>>.

Palestine Info Society (9 May 2008). *Issue Crawler*. 8 Mar 2009 <<http://wiki.IssueCrawler.net/Pisp/ProjectsPage>>.

Pearce-Moses, Richard (2005). "SAA: Glossary of Archival Terminology." *The Society of American Archivists*. 8 Mar 2009 <http://www.archivists.org/glossary/term_details.asp?DefinitionKey=156>.

Planet (2004). *Planet in Internet Archive*. 8 Mar 2009

<http://Web.archive.org/Web/*sr_661nr_30/http://www.planet.nl/planet/show?id*>.

PNINA - "The Official .ps ccTLD Domain Names Registry" (2004). *Palestinian National Naming Authority*. 8 Mar 2009 <<http://www.ps/>>.

Ras, M., and S. van Bussel (2007). "Web Archiving User Survey. Technical Report." *Koninklijke Bibliotheek*.

<www.kb.nl/hrd/dd/dd_projecten/Webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf>

Ras, Marcel, and Barbara Sierman (2006). "Long-term Preservation and Access of the Dutch Web." *Koninklijke Bibliotheek*. 9 Mar 2009.

<www.kb.nl/hrd/dd/dd_projecten/Webarchivering/documenten/IWAW2006_Ras.pdf>

Reiss, Spencer (2 Dec 1998). "Internet in a Box." *Wired in the Internet Archive*. 8 Mar 2009

<<http://Web.archive.org/Web/19981202094728/www.wired.com/wired/4.10/scans.html>>.

Reporters Without Borders (2008). "Internet Enemies". *Reporters Sans Frontières*. 8 Mar 2009

<http://www.rsf.org/article.php3?id_article=26082>.

Rogers, Richard (5 Oct 2008a). "Introduction Digital Methods Initiative." *Digital Methods Initiative*. 8 Mar 2009 <<http://wiki.digitalmethods.net/Dmi/MoreIntro>>.

---. (10 Nov 2008b). "The Demise of the Directory." *Digital Methods Initiative*. 8 Mar 2009

<<http://wiki.digitalmethods.net/Dmi/DemiseDirectory>>.

SIDN - "Website Stats" (2008). *Stichting Domein Registratie Nederland*. 9 Mar 2009

<http://www.sidn.nl/ace.php/p.727,5574,1469347224,Website_stats_2008-11_NL_pdf>.

Sterling, Bruce (1996). "Dead Medium: Internet Archival Issues Part Two." *Dead Media Project*. 8

Mar 2009 <<http://www.deadmedia.org/notes/26/264.html>>.

TBTF - "Disturbing Napier's Bones" (28 Jul 1997). *Tasty Bits from the Technology Front*. 8 Mar 2009

<<http://www.tbtf.com/archive/1997-07-28.html>>.

UNESCO - "UNESCO Charter on the Preservation of the Digital Heritage" (2003). *UNESCO*. 8 Mar 2009 <<http://portal.unesco.org/ci/en/ev.php->

[URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html](http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html)>.

UNESCO - "What Is It? What Does It Do?" (2009). *UNESCO*. 8 Mar 2009
 <http://portal.unesco.org/en/ev.php-URL_ID=3328&URL_DO=DO_TOPIC&URL_SECTION=201.html>.

Villeneuve, Nart (2006). "Google.cn Filtering: How It Works." *OpenNet Initiative*.
 <<http://opennet.net/blog/2006/01/googlecn-filtering-how-it-works>>

Wikipedia - "Bots" (2009). *Wikipedia, the Free Encyclopedia*. 8 Mar 2009
 <http://en.wikipedia.org/wiki/Category:Wikipedia_bots_by_name>

Wikipedia - "TLD" (2009). *Wikipedia, the Free Encyclopedia*. 8 Mar 2009
 <http://en.wikipedia.org/wiki/Country_code_top-level_domain>

Wikipedia - ".us. (2009)" *Wikipedia, the Free Encyclopedia*. 8 Mar 2009
 <<http://en.wikipedia.org/wiki/.us>>.

Wikipedia - "User:Robbot" (2009). *Wikipedia, the Free Encyclopedia*. 8 Mar 2009
 <<http://en.wikipedia.org/wiki/User:Robbot>>.

Zuckerman, Ethan (31 Oct 2008). "The Polyglot Internet." *Ethan Zuckerman*. 8 Mar 2009
 <<http://www.ethanzuckerman.com/blog/the-polyglot-Internet/>>.

Figures References

Figure 1. "tja... national Webs...." (24 July 2008). *Email Correspondence with Geert Lovink*.

Figure 2. Sorry. *ABC Television*. 9 Mar 2009 <<http://www.abc.net.au/tv/sorry.htm>>.

Figure 3. This video is not available in your country (2008). *YouTube.com* <
http://www.youtube.com/index?ytsession=u2Y6A0196wdzmD4G_hY_UtfnmlpH1iUySB7GFYK49tZDNIkhAmOnz_1bKrD6DiSvUqRRfUg7H9hN6ptg_1Vfd2F-QGsmBwCtHJKvD5nnbGGAWbOJH5h71y7ylHUt9JdkGT4RbQwvCXDjNrspAxA0iSEhc0yWHMIsXrXiRK B6lr76zgzW3EAr14UcA_r151MaGeaWiin3kWmaXrgs5S3Rj79ysEuTMCI6ZzTLlT0CTxHBuuUJMEa3k51b5HKExa2ZP645_qz1n3dkm63sY7c_Gg>

Figure 4. 'Video not available in my country' (2008). *YouTube Discussions*.

<http://help.youtube.com/group/youtube-is-sues/browse_thread/thread/9aaa53702203f76b/dbec4cf77b7dff00?lnk=gst&q=%E2%80%98Video+not+available+in+my+country%E2%80%99+#dbec4cf77b7dff00>

Figure 5. Cyberspace I: The Matrix (2008). *The Matrix in Valdosta Museum*.

<<http://www.valdostamuseum.org/hamsmith/MatrixCode.gif>>

Figure 6. Digital Divide Cartogram (2005). *Govcom.org*.

<http://govcom.org/maps/map_set_wsis/GC0_Maps_set_3.0_digitaldivide.pdf>

Figure 7. Cyberspace II: CyberMap Landmarks John December (1994). *Mundi.net*

<http://www.mundi.net/maps/maps_010/johndec_map1.html>

Figure 8. TCP/IP layered model of the Internet infrastructure (2009). *UnixWare 7 Documentation*

<http://uw713doc.sco.com/en/NET_tcpip/tcpN.tcpip_stack.html>

Figure 9. Flickr FAQ Content filters (2009) *Flickr*. <<http://www.flickr.com/help/filters/>>

Figure 10. Blocked photos by Flickr (2008). *Flickr*

<<http://www.flickr.com/photos/leonloes/2989071913/>>

Figure 11. [Official Topic] Filters (2008). *Flickr* <<http://www.flickr.com/help/forum/en-us/35971/page6/#reply226035>>.

Figure 12. Legal requirements, [Official Topic] Filters (2008). *Flickr*.

<<http://www.flickr.com/help/forum/en-us/42597/>>

Figure 13. Censorship circumvention strategy (2007). *Flickr* <<http://www.flickr.com/help/forum/en-us/42597/>>

Figure 14. Cyberspace's fiber-optic cable network (2008). *NRC Handelsblad*

<http://www.nrc.nl/multimedia/archive/00170/270808ECO_glasvezel_170985a.jpg>

Figure 15. Table of RIRs

Figure 16. RIRs IPv4 Whois Map (2007). *Caida* < <http://www.caida.org/research/id-consumption/whois-map/images/whois-20071001.png>>

Figure 17. Mapping the Palestinian Web Space. A Comparison of where .ps sites are hosted and registered (2007). *Information Society in Palestine Project*.

<http://govcom.org/gco_projects/pisp/GC0_Maps_set_ps_3.pdf>

Figure 18. Country Codes of the World (2008). *Bytelevel Research* <

<http://bytelevel.com/map/ccTLD.html>>

Figure 19. The World according to Google I (2008). *Digital Methods Initiative*.

<<http://www.digitalmethods.net/Digitalmethods/TheWebs>>

Figure 20. The World according to Google II (2008). *Digital Methods Initiative*.

<<http://www.digitalmethods.net/Digitalmethods/TheWebs>>

Figure 21. The World according to Google III (2008). *Digital Methods Initiative*.

<<http://www.digitalmethods.net/Digitalmethods/TheWebs>>

Figure 22. The World according to Google IV (2008). *Digital Methods Initiative*.

<<http://www.digitalmethods.net/Digitalmethods/TheWebs>>

Figure 23. The World according to Google V (2008). *Digital Methods Initiative*.

<<http://www.digitalmethods.net/Digitalmethods/TheWebs>>

Figure 24. Internet Archive is creating a library of bits and bytes (1997). *Internet Archive*

<<http://Web.archive.org/Web/19971011064403/http://www.archive.org/index.html>>.

Figure 25. Expert user search strategy with a central position for the search engine, Hölischer and Strube 2000

Figure 26. Tape Jukebox ADIC Scalar 448. *Unylogix.com* <

http://www.unylogix.com/data_storage/tapes_jukebox/adic/images/dlt_scalar458.gif>

Figure 27. "The bits go here. A sample Internet Archive server rack, encompassing a petabyte of storage. A petabyte is 1000 terabytes, and a terabyte is 1000 gigabytes,"

(2009). *The Internet Archive* <<http://www.archive.org/Web/petabox.php>>.

Figure 29. Alexa.com through the Wayback Machine (1997). *Alexa in the Internet Archive*.

<<http://Web.archive.org/Web/19970530104435/http://www.alexa.com/>>.

Figure 28. The physical data center of the Internet Archive, Bibliotheca Alexandrina (2009). *The Internet Archive*. <<http://ia331408.us.archive.org/2/items/bibalex/8598-lg.jpg?cnt=0>>

Figure 30. The first Alexa.com archived page in the Internet Archive (1997). *Alexa in the Internet Archive* <<http://Web.archive.org/Web/19970530104435/http://www.alexa.com/>>

Figure 31. Archive.alexa.com in the Internet Archive (2009) *Archive.alexa.com in the Internet Archive*

<http://web.archive.org/web/*/http://Archive.alexa.com>

Figure 32. Crawl me! Internet Archive Crawler (1997). *The Internet Archive in the Internet Archive*.
<<http://Web.archive.org/Web/19971211124750/www.archive.org/Webmasters.html>>

Figure 33. IIPC inter-actor (members) (2009). *Issue Crawler*. <http://IssueCrawler.net/svg2/issue4.php?id_session=c74b7d9fbd51e8c2273f4953ce618e79idnoId>

Figure 34. IIPC inter-actor (members) - Internet Archive (2009). *Issue Crawler*. <http://IssueCrawler.net/svg2/issue4.php?id_session=c74b7d9fbd51e8c2273f4953ce618e79idnoId>

Figure 35. IIPC inter-actor (members) - European Archive Foundation (2009). *Issue Crawler*.
<http://IssueCrawler.net/svg2/issue4.php?id_session=c74b7d9fbd51e8c2273f4953ce618e79idnoId>

Figure 36. IIPC inter-actor (members) - Royal Library of the Netherlands (2009). *Issue Crawler*.
<http://IssueCrawler.net/svg2/issue4.php?id_session=c74b7d9fbd51e8c2273f4953ce618e79idnoId>

Figure 37. Extensive archiving (shaded area). Some pages are missing (a3, c6) as well as the 'hidden' part of sites (DB, Files), Masanès 2005

Figure 38. Intensive archiving (shaded area). Aims to collect fewer sites but collects deeper content, including potentially parts of the "hidden" Web, Masanès 2005

Figure 39. Table of 'national aspect' criteria (see Appendix C for the extended selection criteria for Web archiving by the KB (in Dutch).

Figure 40. Tools in the KB archiving scheme, KB Technical Aspects (2009). *Koninklijke Bibliotheek*.
<http://www.kb.nl/hrd/dd/dd_projecten/Webarchivering/technisch-en.html>.

Figure 41. Great Chain of Being (2009). *Wikipedia, the Free Encyclopedia*. <http://en.wikipedia.org/wiki/File:Great_Chain_of_Being_2.png>

Figure 42. Dewey Decimal Classification (2009). *ShopBrodArt*.
<<http://www.shopbrodart.com/shop/thumb/p.aspx?p=70&pgid=1799>>

Figure 44. Selection Demise of the Directory (2008). *Digital Methods Initiative* <<http://wiki.digitalmethods.net/Dmi/DemiseDirectory>>

Figure 43. Notification in Wikipedia.org article on Knowledge (2008). *Wikipedia, the Free encyclopedia* <<http://en.wikipedia.org/wiki/Knowledge>>

Appendix A

<i>ccTLD</i>	<i>Country</i>	<i>Foreign registration permitted</i>	<i>Vanity ccTLD (source: Wikipedia TLD 2009)</i>	<i>nr of results in Google.com</i>	<i>nr of results in Region Search</i>
.ac	Ascension Island	yes	no	n/a	n/a
.ad	Andorra	no	yes	580000	297000
.ae	United Arab Emirates	no	no	3770000	3650000
.af	Afghanistan	no	no	168000	162000
.ag	Antigua and Barbuda	yes	yes	684000	662000
.ai	Anguilla	no	no	66400	66400
.al	Albania	no	no	416000	2350000
.am	Armenia	yes	yes	3370000	3850000
.an	Netherlands Antilles	no	no	31900	32200
.ao	Angola	no	no	210000	493000
.aq	Antarctica	no	no	73	73
.ar	Argentina	no	no	108000000	107000000
.as	American Samoa	yes	yes	819000	28
.at	Austria	yes	no	198000000	196000000
.au	Australia	no	no	255000000	254000000
.aw	Aruba	no	no	17800	17700
.ax	Åland Islands	no	no	n/a	n/a
.az	Azerbaijan	no	no	2970000	25400000
.ba	Bosnia and Herzegovina	no	no	9420000	9530000
.bb	Barbados	no	no	234000	236000
.bd	Bangladesh	no	no	n/a	n/a
.be	Belgium	yes	yes	n/a	n/a
.bf	Burkina Faso	no	no	133000	132000
.bg	Bulgaria	no	no	64200000	65700000
.bh	Bahrain	no	no	546000	480000
.bi	Burundi	yes	no	53300	53100
.bj	Benin	no	no	17900	17900
.bm	Bermuda	no	no	222000	680000
.bn	Brunei	no	no	893000	1010000
.bo	Bolivia	yes	no	3150000	3010000
.br	Brazil	yes	no	399000000	405000000
.bs	Bahamas	yes	no	124000	124000
.bt	Bhutan	no	no	77100	77400
.bv	Bouvet Island (not in use; no registrations)	no	no	n/a	n/a
.bw	Botswana	no	no	190000	562000
.by	Belarus	no	no	10400000	10900000
.bz	Belize	yes	no	2460000	209000
.ca	Canada	no	no	288000000	284000000
.cc	Cocos (Keeling) Islands	yes	yes	137000000	1
.cd	Democratic Republic of the Congo (formerly .zr)	yes	yes	514000	76
.cf	Central African Repub-	no	no	11	11

	lic				
.cg	Republic of the Congo	yes	no	84	84
.ch	Switzerland	yes	no	187000000	185000000
.ci	Côte d'Ivoire (Ivory Coast)	yes	no	288000	317000
.ck	Cook Islands	yes	no	132000	131000
.cl	Chile	no	no	56500000	56200000
.cm	Cameroon	no	no	238000	993000
.cn	People's Republic of China	yes	no	1110000000	80000000
.co	Colombia	no	no	26300000	26100000
.cr	Costa Rica	no	no	n/a	n/a
.cu	Cuba	no	no	n/a	n/a
.cv	Cape Verde	no	no	n/a	n/a
.cx	Christmas Island	yes	no	n/a	n/a
.cy	Cyprus	no	no	2380000	2360000
.cz	Czech Republic	no	no	313000000	309000000
.de	Germany	no	no	1480000000	1530000000
.dj	Djibouti	yes	yes	481000	121000
.dk	Denmark	yes	no	205000000	206000000
.dm	Dominica	no	no	414000	420000
.do	Dominican Republic	no	no	2090000	2110000
.dz	Algeria	no	no	383000	141000
.ec	Ecuador	yes	no	7800000	7450000
.ee	Estonia	no	no	136000000	138000000
.eg	Egypt	no	no	3890000	3740000
.eh	Western Sahara (not assigned; no DNS)	no	no	n/a	n/a
.er	Eritrea	no	no	32	32
.es	Spain	yes	no	376000000	378000000
.et	Ethiopia	no	no	96500	185000
.eu	European Union (code "exceptionally reserved" by ISO 3166-1)	no	no	n/a	n/a
.fi	Finland	no	no	152000000	148000000
.fj	Fiji	yes	no	253000	253000
.fk	Falkland Islands	no	no	68	68
.fm	Federated States of Micronesia	yes	yes	21600000	74
.fo	Faroe Islands	no	no	1280000	1280000
.fr	France	no	no	576000000	571000000
.fx				n/a	n/a
.ga	Gabon	no	no	13200	25200
.gb	United Kingdom (Reserved domain by IANA; deprecated)	no	no	n/a	n/a
.gd	Grenada	yes	no	85400	85500
.ge	Georgia	no	no	7240000	9630000
.gf	French Guiana	no	no	53	53
.gg	Guernsey	no	yes	n/a	n/a
.gh	Ghana	no	no	425000	496000

.gi	Gibraltar	no	no	187000	188000
.gl	Greenland	yes	no	627000	1480000
.gm	Gambia	no	no	164000	164000
.gn	Guinea	no	no	37	37
.gp	Guadeloupe	no	no	184000	1300000
.gq	Equatorial Guinea	no	no	12	13
.gr	Greece	yes	no	118000000	118000000
.gs	South Georgia and the South Sandwich Islands	yes	no	851000	863000
.gt	Guatemala	no	no	2210000	2190000
.gu	Guam	no	no	47	47
.gw	Guinea-Bissau	no	no	25	25
.gy	Guyana	no	no	92200	92700
.hk	Hong Kong	yes	no	54200000	52600000
.hm	Heard Island and McDonald Islands	yes	no	227000	108000
.hn	Honduras	yes	no	1470000	1950000
.hr	Croatia	no	no	55400000	56300000
.ht	Haiti	no	no	59000	59100
.hu	Hungary	yes	no	247000000	247000000
.id	Indonesia	no	no	54500000	55500000
.ie	Ireland	no	no	35000000	33300000
.il	Israel	yes	no	281000000	281000000
.im	Isle of Man	yes	yes	627000	627000
.in	India	yes	yes	97300000	93900000
.io	British Indian Ocean Territory	yes	no	250000	249000
.iq	Iraq	no	no	74	74
.ir	Iran	yes	no	30900000	32100000
.is	Iceland	yes	no	27100000	28300000
.it	Italy	no	yes	574000000	570000000
.je	Jersey	no	yes	n/a	n/a
.jm	Jamaica	no	no	251000	252000
.jo	Jordan	no	no	847000	823000
.jp	Japan	no	no	1750000000	1720000000
.ke	Kenya	no	no	537000	532000
.kg	Kyrgyzstan	no	no	1130000	1130000
.kh	Cambodia	no	no	214000	493000
.ki	Kiribati	no	no	78700	78700
.km	Comoros	no	no	21	21
.kn	Saint Kitts and Nevis	no	no	81	81
.kp	North Korea	no	no	3	3
.kr	South Korea	no	no	391000000	383000000
.kw	Kuwait	no	no	1090000	1150000
.ky	Cayman Islands	no	no	197000	197000
.kz	Kazakhstan	yes	no	5340000	5620000
.la	Laos	yes	yes	4230000	290000
.lb	Lebanon	no	no	942000	1860000
.lc	Saint Lucia	no	no	16900	16900
.li	Liechtenstein	yes	yes	1960000	1940000

.lk	Sri Lanka	no	no	937000	905000
.lr	Liberia	no	no	95	73
.ls	Lesotho	yes	no	240000	185000
.lt	Lithuania	no	no	106000000	106000000
.lu	Luxembourg	no	no	5160000	5000000
.lv	Latvia	yes	yes	56900000	58700000
.ly	Libya	yes	no	439000	245000
.ma	Morocco	no	no	3180000	3050000
.mc	Monaco	no	no	224000	225000
.md	Moldova	yes	yes	4970000	5000000
.me	Montenegro	no	yes	n/a	n/a
.mg	Madagascar	no	no	241000	571000
.mh	Marshall Islands	no	no	n/a	n/a
.mk	Republic of Macedonia	no	no	3070000	4800000
.ml	Mali	no	no	345000	336000
.mm	Myanmar	no	no	89200	89300
.mn	Mongolia	yes	no	1160000	13700000
.mo	Macau	no	no	1240000	1120000
.mp	Northern Mariana Islands	yes	no	12	12
.mq	Martinique	no	no	61	61
.mr	Mauritania	no	no	84200	84800
.ms	Montserrat	yes	yes	1120000	30
.mt	Malta	no	no	1060000	1120000
.mu	Mauritius	yes	yes	541000	541000
.mv	Maldives	no	no	627000	1210000
.mw	Malawi	yes	no	86900	86900
.mx	Mexico	yes	no	202000000	201000000
.my	Malaysia	no	no	20600000	6610000
.mz	Mozambique	no	no	574000	1470000
.na	Namibia	yes	no	424000	413000
.nc	New Caledonia	no	no	231000	396000
.ne	Niger	no	no	85	85
.nf	Norfolk Island	yes	no	172000	221000
.ng	Nigeria	no	no	406000	388000
.ni	Nicaragua	no	no	3140000	3100000
.nl	Netherlands	yes	no	421000000	8720000
.no	Norway	no	no	220000000	222000000
.np	Nepal	no	no	1710000	609000
.nr	Nauru	yes	no	n/a	n/a
.nu	Niue	yes	yes	35300000	12
.nz	New Zealand	yes	no	61700000	60300000
.om	Oman	no	no	414000	408000
.pa	Panama	no	no	2390000	2250000
.pe	Peru	no	no	9390000	8810000
.pf	French Polynesia	no	no	120000	155000
.pg	Papua New Guinea	no	no	64200	64200
.ph	Philippines	yes	no	12000000	11800000
.pk	Pakistan	yes	no	3570000	3940000
.pl	Poland	yes	no	544000000	545000000
.pm	Saint Pierre and	no	no	n/a	n/a

	Miquelon				
.pn	Pitcairn Islands	yes	no	109000	108000
.pr	Puerto Rico	yes	no	599000	684000
.ps	Palestine	yes	no	975000	976000
.pt	Portugal	yes	no	90900000	92800000
.pw	Palau	no	no	1	1
.py	Paraguay	no	no	1840000	1980000
.qa	Qatar	no	no	1050000	1030000
.re	Réunion	no	no	160000	160000
.ro	Romania	yes	no	237000000	230000000
.rs	Serbia	yes	no	n/a	n/a
.ru	Russia	yes	no	878000000	875000000
.rw	Rwanda	no	no	194000	525000
.sa	Saudi Arabia	no	no	7200000	7600000
.sb	Solomon Islands	yes	no	109000	269000
.sc	Seychelles	yes	yes	336000	36300
.sd	Sudan	no	no	55300	48100
.se	Sweden	yes	no	242000000	241000000
.sg	Singapore	no	no	16000000	15100000
.sh	Saint Helena	yes	no	659000	654000
.si	Slovenia	no	no	44500000	45700000
.sj	Svalbard and Jan Mayen islands (not in use; no registrations)	no	no	n/a	n/a
.sk	Slovakia	no	no	177000000	176000000
.sl	Sierra Leone	no	no	58	58
.sm	San Marino	yes	no	1340000	473000
.sn	Senegal	no	no	277000	268000
.so	Somalia (down, still is delegated to Monolith [ml.org] Philadelphia, an entity defunct since end-1998)	yes	no	2	2
.sr	Suriname	yes	no	249000	13400
.st	São Tomé and Príncipe	yes	yes	3080000	2700000
.su	Soviet Union (depre- cated; being phased out; code "transitionally reserved" by ISO 3166-1)	no	no	11300000	11300000
.sv	El Salvador	no	no	2370000	2360000
.sy	Syria	yes	no	502000	502000
.sz	Swaziland	yes	no	54800	54700
.tc	Turks and Caicos Is- lands	yes	no	1740000	1660000
.td	Chad	no	no	2	2
.tf	French Southern Terri- tories	no	no	118000	118000
.tg	Togo	yes	no	17100	17100
.th	Thailand	yes	no	96200000	98200000
.tj	Tajikistan	yes	no	380000	636000
.tk	Tokelau	yes	no	3810000	1

.tl	East Timor (formerly .tp)	yes	no	n/a	n/a
.tm	Turkmenistan	yes	no	18	18
.tn	Tunisia	no	no	808000	800000
.to	Tonga	yes	yes	20000000	19400000
.tp	East Timor (depre- cated)	no	no	177000	n/a
.tr	Turkey	no	no	203000000	208000000
.tt	Trinidad and Tobago	yes	no	948000	924000
.tv	Tuvalu	yes	yes	68000000	9
.tw	Taiwan	yes	no	195000000	193000000
.tz	Tanzania	no	no	262000	1400000
.ua	Ukraine	no	no	153000000	156000000
.ug	Uganda	yes	no	338000	336000
.uk	United Kingdom (code "exceptionally re- served" by ISO 3166- 1) (see also .gb)	no	no	780000000	789000000
.um				n/a	n/a
.us	United States	yes	no	169000000	164000000
.uy	Uruguay	no	no	4370000	4030000
.uz	Uzbekistan	no	no	3470000	3610000
.va	Vatican City	no	no	41	40
.vc	Saint Vincent and the Grenadines	yes	no	1160000	472000
.ve	Venezuela	no	no	12400000	3000000
.vg	British Virgin Islands	yes	yes	733000	727000
.vi	United States Virgin Islands	no	no	82600	82500
.vn	Vietnam	no	no	9980000	87000000
.vu	Vanuatu	yes	yes	527000	n/a
.wf	Wallis and Futuna	no	no	5	n/a
.ws	Samoa (formerly Western Samoa)	yes	yes	32400000	n/a
.ye	Yemen	no	no	49500	n/a
.yt	Mayotte	no	no	1	n/a
.za	South Africa	yes	no	27400000	n/a
.zm	Zambia	no	no	138000	n/a
.zw	Zimbabwe	no	no	295000	n/a

Appendix B

<i>IIPC member</i>	<i>URL</i>
Biblioteca de Catalunya (Library of Catalonia)	http://www.bnc.cat
Biblioteca Nazionale Centrale di Firenze (National Library of Italy, Florence)	http://www.bncf.firenze.sbn.it/
Biblioteka Narodowa (National Library of Poland)	http://bn.org.pl/
Bibliothèque et Archives nationales du Québec (BAnQ)	http://www.banq.qc.ca/
Bibliothèque nationale de France (National Library of France)	http://www.bnf.fr
British Library (U.K.)	http://www.bl.uk
California Digital Library (U.S.)	http://www.cdlib.org
Centre for Global eHealth Innovation, WebCite® Internet Citations Archiving Project (Canada)	http://www.webcitation.org
Deutsche Nationalbibliothek (German National Library)	http://www.d-nb.de
European Archive Foundation	http://europarchive.org
Hanzo Archives Ltd. (U.K.)	http://www.hanzoarchives.com
Ina (Institut National de l'Audiovisuel) (France)	http://www.ina.fr/
Internet Archive (U.S.)	http://www.archive.org
Internet Preservation Consortium (IIPC)	http://netpreserve.org
Jewish National and University Library (Israel)	http://www.jnul.huji.ac.il/IA/ArchivedSites/IA/firstpage.html
Kansalliskirjasto (National Library, Finland)	http://www.lib.helsinki.fi
Koninklijke Bibliotheek (National Library of the Netherlands)	http://www.kb.nl
Kungl. biblioteket, (National Library of Sweden)	http://www.kb.se
Landsbokasafn Islands – Haskolabokasafn (National and University Library of Iceland)	http://www.bok.hi.is
Latvijas Nacionālā bibliotēka (National Library of Latvia)	http://www.lnb.lv
Library and Archives Canada	http://www.collectionscanada.ca
Library of Congress (U.S.)	http://www.loc.gov/webcapture
Nacionalna i sveučilišna knjižnica u Zagrebu (National and University Library in Zagreb, Croatia)	http://www.nsk.hr/digarhiv
Narodna in univerzitetna knjižnica (National and University Library, Slovenia)	http://www.nuk.uni-lj.si
Národní knihovna České republiky (National Library of the Czech Republic)	http://www.nkp.cz
Nasjonalbiblioteket (National Library of Norway)	http://www.nb.no
National Archives (U.K.)	http://www.nationalarchives.gov.uk
National Diet Library, Japan	http://www.ndl.go.jp/
National Library Board, Singapore	http://www.nlb.gov.sg
National Library of Australia	http://www.nla.gov.au
National Library of China	http://www.nlc.gov.cn/en/indexen.htm
National Library of Korea	http://www.oasis.go.kr
National Library of New Zealand	http://www.natlib.govt.nz
National Library of Scotland	http://www.nls.uk
Netarchive.dk	

(Royal Library and the State and University Library, Aarhus)	http://www.netarchive.dk
Österreichische Nationalbibliothek	
(Austrian National Library)	http://www.onb.ac.at/
Schweizerische Nationalbibliothek	
(Swiss National Library)	http://www.nb.admin.ch
United States Government Printing Office	http://www.gpo.gov/projects/fdsys.htm
University of North Texas Libraries (U.S.)	http://www.library.unt.edu/
Virtual Knowledge Studio –	
Royal Netherlands Academy for Arts and Sciences	http://www.virtualknowledgestudio.nl/index-temp.php
	source: http://netpreserve.org/about/members.php

Appendix C

KB selectiecriteria webarchivering*

1	Domein	.nl en andere in Nederland geregistreerde domeinen	
2	Nationale aspect		
	A	Website in Nederlands en geregistreerd in Nederland	
	B	Website in andere taal, geregistreerd in Nederland	
	C	Website in Nederlands, geregisterd in ander land	
	D	Website in andere taal, geregistreerd in ander land, onderwerp gericht op Nederland	
3	Content	Bronnen met culturele, wetenschappelijke waarde. Websites met betrekking tot Nederlandse taal, cultuur en samenleving (collectiebeleid KB) + output van de overheid. Ook web 2.0 toepassingen als weblogs en innovatie en trends op het web.	
4	Protocol	http	
5	IPR	Via opt-out. Auteur/beherende organisatie/persoon moet bekend en traceerbaar zijn	
6	Toegang	Alleen openbaar toegankelijke sites. Mogelijk ook dieper wanneer individuele afspraken gemaakt kunnen worden. Respecteren van robots.txt	
7	Formaat	Alle gebruikelijke bestandsformaten die met standaard browsers en standaard plug-ins geïnterpreteerd kunnen worden. Technische grenzen zijn afhankelijk van harvester. Digitale duurzaamheid is een afhankelijke ten aanzien van presentatie gearchiveerde website	
8	brontype	Website behandelen als en geheel, geen selecties maken van delen van websites. (geen losse componenten)	
	Niet verzamelen	Games, portals, online nieuws (?), webcams, msn, datasets, bulletin boards, intranets, RTV programma's. Websites die al verzameld worden in het kader van andere webarchiveringsprojecten in Nederland. Daarmee moet samenwerking aangegaan worden.	

* Buiten de selectie vallen onderstaande categorieën. Deze websites worden door de genoemde instituten verzameld of zullen verzameld worden in de (nabije) toekomst:

Rotterdamse websites en websites over Rotterdam (Gemeentearchief R'dam)

Websites van politieke partijen (Archipol, Groningen)

Websites van (voornamelijk publieke) omroepen (Ned. Inst. voor Beeld en Geluid)

Websites mbt het vakgebied Sinologie (UBL)