Archiving the web

Voerman, Gerrit; Keyzer, Andreas; den Hollander, Franciscus; Druiven, Hendricus

Published in:
European Political Science

DOI:
10.1057/eps.2002.51

Publication date:
2002

Citation for published version (APA):
Voerman, G., Keyzer, A., den Hollander, F., & Druiven, H. (2002). Archiving the web: political party web sites in the Netherlands. *European Political Science*, *2*(1). https://doi.org/10.1057/eps.2002.51

ARCHIVING THE WEB:

political party web sites in the netherlands

Gerrit Voerman, André Keyzer Frank den Hollander and Henk Druiven

"*Pantha rhei,*" was one of the tenets of the Greek philosopher Heraclitus. This observation on life in Greece in the 6th century BC is certainly relevant to the World Wide Web some 2500 years later. On the Web everything is in a state of flux and is subject to continuous change. The Web is expanding at an enormous pace; millions of new pages appear every month. The number of sites – the building blocks of the Web – is supposed to increase with millions each year and has amounted to 162 million at July 2002 (Internet Software Consortium, www.isc.org/ds/WWW-200207/index.htlm). This incredible expansion is taking place despite the fact that at the same time as new sites are created, many disappear: the average lifes pan of a site is estimated at seventy-five days. Sites are often provided at only one location. If, for whatever reason, a supplier decides to discontinue a site, it is lost forever. In this dynamic process of rise and fall most existing sites are not static either. They keep changing all the time: a few seconds after visiting a site, it may have changed because the supplier added information or a visitor leaves a message.

Given this ongoing transformation of the Web, one might have expected that earlier manifestations of sites to have been preserved. For a long time after the advent of the Web in the early 1990s, however, this was not the case. On the contrary, the "digital memory" was allowed to disappear completely. American historians attending a congress in 1998 noteded that, "the Internet threatens to cause a gap in historiography" (*De Volkskrant*, 30 January 1999). Only recently has a start been made to store Websites and make them accessible (Casey, 1998). This does not alter the fact, however, that a substantial part of our digital inheritance has already been lost. Future researchers looking into the developments of the "virtual" world of the Web, the evolution of the medium and its relationship to the "real" world, will have to face lack of source material.

This article will focus on a project in the Netherlands, where the Web sites of political parties are being archived by the *Documentatiecentrum Nederlandse Politieke Partijen* (Documentation Centre for Dutch Political Parties (DNPP) of the University of Groningen). First, however, several important international initiatives will be briefly described. Then we will focus on the growing importance of Web sites for political parties. After that, the Dutch project (called Archipol: ARCHIving POLitical parties' Web sites; www.archipol.nl) will be dealt with. The article ends with a discussion of the significance of archived Web sites for political scientists.

International initiatives to archive the Web

People started to think about downloading and storing Web sites, from the mid-1990 (Arms *et al.* 2001). One of the pioneers was the National Library of Australia, which in June 1996 set up the so-called Pandora project (Preserving and Accessing Networked DOcumentary Resources in Australia). Within the scope of Pandora, Australian on-line publications considered important are being archived, including some Websites of Australian political parties (http://pandora.nla.gov.au/index.html). A few months later, in September 1996, the Royal Library of Sweden started The

Kulturarw[1] Project. The purpose of this project is to archive the Swedish part of the Internet, that is, all URLs with the extension dot.*se*. Up to now, several snapshots have been produced, involving the storage of around 126,000 Web sites in total. The digital archive has not yet been made accessible to the general public (Arvidson and Lettenström, 1998; http://www.kb.se/kw3/ENG/Description.htm). Meanwhile the Swedish project has been copied by the National Library in Finland. The first round of harvesting of the Finnish Web was completed in June 2002.

The most ambitious project, however, is one that has been developed by the American computer programmer Brewster Kahle. He set up the San Francisco-based Internet Archive (www.archive.org), which, since the summer of 1996, has been busy archiving the Internet – as its name suggests – from newsgroups to homepages. In order to achieve this, Web crawling robots are used: programmes that find sites by means of external links of other sites, and subsequently download them all, in their entirety. In this way a random picture of the Internet is provided. By October 2001, 10 billion unspecified Webpages had been gathered, as well as around 16 million news items - mainly HTLM files (Kahle, 1997; Cunningham, 1997; Editor's interview, 2002). In the autumn of 2001, this collection became available on the internet through the so-called 'Wayback Machine'.

The Internet Archive has a few drawbacks, however. First of all, the method used is fairly rough-and-ready, and is aimed at quantity: it involves storing as many sites as possible. Within this bulk collection strategy it is difficult to guarantee the quality of an archived site. Furthermore, the archive has not been made accessible with respect to content. That is, to find a site in the archive, one has to know the exact URL.

In contrast to the American Internet Archive, the DNPP archiving project has opted for a selective approach, aimed at a specific, limited category of Web sites. Since political parties first appeared on the Web, most of them have completely restyled their sites three or four times. Almost nothing has remained of their older versions (except a few which were included in the Internet Archive, as became clear when the Wayback Machine was launched). Therefore, it is not possible to retrace the initial steps taken by Dutch political parties on the Web. In 2000 the DNPP, together with the University Library of the University of Groningen, started to preserve parties' Web sites alongside the library's printed collection.[2] The project is aimed at the Websites of the political parties represented in the Dutch Parliament and those of their subsidiary organisations, especially youth organisations, but also at parties that are not represented in Parliament.

Why archive political parties' Websites?

Dutch political parties could be found on the Web at a fairly early stage. In January 1994 *GroenLinks* (GreenLeft) – an environmentalist party – was the first party represented in parliament to start a Website (Ward and Voerman, 2000). Then, more or less in a movement from left to right across the political spectrum, the other parties followed suit: *Partij van de Arbeid* (PvdA, Labour Party) in November 1994, the left-liberal *Democraten 66* (D66, Democrats 66) and *Christen Democratisch Appèl* (CDA, Christian Democratic Party) about the middle of 1995, and the right-liberal *Volkspartij voor Vrijheid en Democratie* (VVD, People's Party for Freedom and Democracy) in the spring of 1997. The digital début of the *Socialistische Partij* (SP, Socialist Party) does not entirely fit this picture: it made its first appearance on the Internet in autumn 1997. The last party represented in parliament to get on line was the orthodox-protestant *Staatkundig-Gereformeerde Partij* (SGP; Political Reformed Party), in the autumn of 2000 (Voerman and DeGraaf, 1998).

By the eve of the general election of May 1998, all parties – with the exception of the SGP – had a Web site. During the election campaign Web sites were considered not to be very important, because they did not attract many visitors. At most, an estimated 100,000 visited a party Web site during the month preceding election day (Voerman, 2000: 206). Four years later, however, in the spring of 2002, this number had increased spectacularly. The extraordinary situation in which the campaign took place – the rise of the electorally very succesful right-populist politician, Pim Fortuyn, and the fact that he was murdered one-and-a half-weeks before the elections – certainly exerted an influence. As most parties stopped campaigning after the assassination, the Internet became important to citizens in search of information about election programmes, candidates and the latest events. In the last ten days of the campaign the Web sites of the parties represented in Parliament and the site of the Pim Fortuyn List were visited more then a million times. In addition, the most prominent on-line voting guide, *Stemwijzer,* answered more than two million questions – compared to some 6,000 in 1998.

It is clear from these figures that, in a quantative terms, the Web as a channel of communication has grown considerably in importance. Web sites have played a major role in the provision of information to voters and members for some time now – not only during but also outside the electoral season. Electoral programmes, press releases, candidates' backgrounds, articles from party periodicals: not only do they appear in print, but they can all usually be found on the sites, too. Parties are also becoming more inclined to provide certain information only in digital form, not only because it is much faster to do it that way, but also because it is much cheaper. The changing content of members' magazines bears witness to this. These periodicals are beginning to turn away from the rather dry organisational information (speaking engagements, the agendas of party meetings, etc.) they have hitherto provided. D66, for example, no longer prints documents for its party congresses in its members' magazine but furnishes them through its web site. But the evolution of party sites into exclusive publication outlets for certain categories of information, is not only a result of this shift of emphasis towards digitally provided information. It has also come about because parties work with special Internet editors who write articles exclusively for the sites – particularly during election campaigns.

Parties have become increasingly interesting in sites for other reasons besides the opportunities the Web gives them furnish easily updated information cheaply and quickly. Although the majority of sites are still primarily intended to provide party information – and thus are rather "top-down" in style – the interactive possibilities can be expected to gain more weight in the future. In the United States, for example, Web sites are used to recruit campaign workers and drum up financial support. Further, sites will also play a larger role in the opinion-forming process within parties. Most parties are currently having considerable difficulties retaining members and getting them to be active. The new forms of digital participation which have recently become available offer them novel ways of involving people, particularly young people. In the Netherlands, parties are currently experimenting with the possibilities, with the CDA taking the lead. The Christian Democrats used their site to generate ideas about their programme for the 2002 parliamentary elections. Thousands of visitors to the site participated. The number of useful ideas was not particularly large, but the party certainly got an idea of what was important to its supporters. In addition, an internal CDA network was set up to which party members had exclusive access by means of their membership numbers. The intention is to integrate this digital department into the party's 'conventional' organisational structures.

Sites are thus playing an ever greater role within parties, not only in information provision but also in the provision of more opportunities for participation. Besides those sites that are formally the direct responsibility of the party leadership, sites have also been started by certain sections of parties (for example party commissions), related organisations (youth divisions, women) and parliamentary candidates. Occasionally there are also sites for special categories of members, such as the *GroenLinks* executive site, primarily intended for the district board-members, councillors, members of Parliament, legislators and mayors beloning to the party. Besides these more or less 'official' sites, the past few years have also seen the emergence of 'officious' sites belonging to ad hoc pressure groups within the parties. Within D66, for example, groups of dissatisfied members have several times tried to influence the party line with the help of a web site – and not unsuccessfully. Such means were also employed by opponents of the executive committee of the Pim Fortuyn List. Use is also made of web sites during internal elections. During the elections for a new leader of the PvdA in the spring of 2001, each of the four candidates had their own Web site. Thus, sites are beginning to play a role in internal decision-making process within the parties, too.

Archiving web sites

The developments outlined above were an important reason behind the Documentation Centre's decision to archive the party sites. On the one hand, not recording the digital communications of the parties would have led to an ever-increasing hiatus in the collection of party publications. The chances of losing information published only on parties' Web sites, would otherwise have been rather great. Besides librarians' considerations of this kind, Web sites should also be preserved because of their increasing function as means of communication, mobilisation, participation and pressure. The study of political parties and the ways they communicate internally and externally would be more difficult without archived sites.

As mentioned, the DNPP began to archive parties' sites in 2000. This section will briefly explain the archiving process. With regard to the acquisition of the sites, we decided to download them ourselves, and not to request the webmasters of the parties to provide the data and later mutations. This would have made the project too dependent on the co-operation of the parties. Further, more and more sites are being constructed using database systems. Such sites have no fixed content, their pages being generated on the fly, at the moment when the visitor actually requests them. The provision of such 'dynamic' sites in full, including the tree structure, is very difficult. When archiving, a snapshot is preserved, without the interactive functionality. We have written our own programme for the downloading process, but we also make use of the off-line browser HTTrack. Working with two programmes enables us to check for completeness. The advantage of HTTrack is that new versions are constantly being developed to keep pace with the technological advances of the web. This programme is also being used by the archiving project of the Library of Congress, known as Minerva (Mapping the INternet: the Electronic Resources Virtual Archive; http://www.loc.gov/minerva/).

Archiving, that is, moving the downloaded site to the archive, is virtually automatic. All text, image and sound files are included. Incidentally, it is not always possible to store all the parts of a Web site in an archive. This is because parts of sites are inaccessible to off-line browsers, usually because they are contained in robots.txt files. After consultations with the relevant webmaster, it is often possible to be exempted from this ‘robot exclusion’. Only a snapshot can be made of the interactive, dynamic aspects of a Web site, such as reply forms, a chatboxes and search engines: the interactivity no longer works in an archive. It is usually very difficult to archive parts of sites where use is made of new Web technology such as JavaScript or Flash (a Macromedia application that combines text, animations, illustrations, sound and interaction). Sites where the links to subsequent pages are included in a Flash animation are particularly difficult to process.

When archiving a site, all files which are identical to those in a previous download already in the archive are removed. This process of removing duplicates also adapts all the internal links within a site. They are from then on needed only to link to the Web archive and no longer to the original site on the Web. With external links, that is, directions to web sites outside the site in question, a warning appears informing you that you are leaving the Web archive. An overview of all the necessary changes is stored in a meta-data file. This file also includes the source (URL), the date of the download, and information about the size of the archived site.

The archived sites are stored on our own document server. Future storage may well prove to be a problem. In particular, the short lifespan of hardware and software will lead to problems with regard to the maintenance and accessibility of sites. When hardware, programmes and the text storage formats of audio and video have become obsolete, it should still be possible to consult the sites stored in the archive. However, digital archives will have to be transferred and converted to a new generation of information carriers and software and hardware systems. Whether this migration can be done without (slightly) damaging the authenticity and integrity of digital documents remains to be seen, bearing in mind the current state of technology (Mackenzie Owen, 1998: 24-5; Hodge and Carroll, 1999: 7-8).

In addition to technical problems, the archiving of web sites can also come up against legal snags. By definition, the archiving of sites is copying. Any part of a site – photos, audio and video material, articles, or the design of the site itself - may be protected by copyright When copying a web site and making it available in an archive, these rights have to be taken into account. This is why the project is working with a closed archive for the time being. Only authorised users can gain on-line access to the actual archive. Discussions are being held with the boards of the political parties about making the archive generally accessible, and the related copyright issues.

Possibilities for the researcher

The archive can be accessed in a number of ways. Within the archive, it is possible to search by party (affiliated organisation, candidate, etc.) and date of archiving. The size of every archived site is displayed in bytes, words, files and HTML files. It is also possible to check which parts of a site are new with regard to previous versions stored in the archive. Every archived version also has its own site map, which provides an overview of the site’s set-up.

It is also possible to make a limited analysis of the contents of the Web site versions stored. The archived sites can all be searched by word using the search engine developed within the framework of the project. Thus it is possible to check which words are unique to a site, or used the most, or how often a word has been used over the course of time in the various archived versions of a Web site. These tools enable the researcher to study the development of the standpoints and views of the various parties from a content point of view much more easily than is possible with printed material. Further, the tools simplify the study of the political rhetoric of parties, a field that is still rather neglected partly because of the methodological difficulties invoveld.

The archived Web sites are also ideal for the study of the position and function of the web site within the party organisation. Is the site organised in a 'top-down' way, mainly intended to disseminate information, or 'bottom-up' fashion, in order to get reactions? Is it directed towards the mobilisation of sources of assistance (members, finance)? What is its function during election campaigns? Is it controlled by the party leadership or is it independent? Is the site the instrument of an internal pressure group. In order to determine the formal and informal status and functions of a party site, is it necessary to consult its sources. The navigation system within the archive not only enables diachronic research (the development of a certain site over time) but also synchronic analysis (the comparison of the sites of different parties during a given time period).

Conclusion

Web sites enable parties or candidates to present themselves to a wider audience, without the filter of journalism, in a way that they themselves want. The identities of the parties, or the images they want to present of themselves, can be read between the lines. The activist, extra-parliamentary and sometimes populist, attitude of the SP, for example, is expressed on its site by the many different ways perceived injustices are reacted to and reported to the party office by e-mail. The SGP site reflects the character of its party in a completely different way, namely by not being on line on Sundays. Thus, in different ways, sites say something about the political parties.

This all means that it is vitally important to preserve the digital sources for future scientific research into political parties, and thus that there is every reason to archive parties' Web-based materials alongside their printed publications. This will be of benefit to researchers working in various disciplines (historians, sociologists, political scientists, communication specialists) as wel as to journalists. Any comments on this article and the project described are welcome and may be addressed to info@archipol.nl

[1] See A. Arvidson and F. Lettenström, "The Kulturarw³ Project - The Swedish Royal Web Archive", in: *The Electronic Library*, 16 (1998), 2 (April), 105-108. See also www.Kulturarw3.kb.se/html/projectdescription.html.

[2] This archiving project – called Archipol – is jointly funded by the University of Groningen and the steering committee *Innovatie Wetenschappelijke Informatievoorziening* (Innovation of Scientific Information Provision). SURF NOEMEN!!