The Archives of the Present 2. Dezember, 16.30
Archivo di Stato di Milano, Via Senato 10

Archiving the Web Sites of Political Parties in Germany
A Joint Project of the Archives of Political Foundations Funded by the DFG

Rudolf Schmitz
Archiv der sozialen Demokratie / FES

For the purpose of this project, which has been funded by the German Research Foundation, the archives of five political foundations in Germany have joined forces; these archives being the Archive of Social Democracy of the Friedrich Ebert Foundation, the Archive for Christian Democratic Policy of the Konrad Adenauer Foundation, the Archive for Christian Social Policy of the Hanns Seidel Foundation, the Archive of Liberalism of the Friedrich Naumann Foundation and the Archive 'Grünes Gedächtnis' of the Heinrich Böll Foundation.
In the course of the project, which is scheduled to take two years, we envisage not only creating new Internet archives but also developing exemplary procedures that can be adapted by others.
In the process of optimising the electronic backup of political parties' Internet presence, all archives can rely on the long-term experience of the archive of the Friedrich Ebert Foundation, which heads the project.

As early as 1999, the Archive of Social Democracy decided to tackle the challenge of archiving the Internet presence of the Social Democratic party in order to save this new source category permanently and make it available for research.
In preparation for the DFG project, which started last September, the different archives succeeded in finding a common approach for the recording or –as we call it- the mirroring of websites and the presentation of the archived websites. These common approaches, as well as the similarity of the task, were the real basis for the cooperation between the archives.

Before I present the Internet archive to you, please allow me to say a few words on three important aspects of this project: the task, the process of mirroring and the form of presentation.

First of all, the task: Archiving the Internet presence of the Social Democratic party can only mean archiving the websites of the statutory committees, groupings and initiatives of the SPD. This is also true for their parliamentary groups.
Therefore, it is not our intention -and there is no point in doing so- to completely document the SPD’s appearance on the web with its infinite discussions on

programmes and people held in innumerable forums and chats. This would be an arbitrary undertaking. Even including informal groupings of statutory associations means accepting a certain degree of arbitrariness, since it cannot be guaranteed that all this groupings can be found at all. Essentially, the project will be limited to those websites that have their origin in the SPD. This often demands making strict distinctions.
For example, if the website of a Member of the Bundestag contains a link to a newspaper that published an interview with him, this interview will not be included in the project. However, the archived website still provides the user with the information that such an interview exists, when it was held and where to find it.

A similar approach is taken towards the streaming files offered by the administration of the Bundestag, to which numerous Members of the Bundestag set links on their websites. (The administration puts over 500 hours of video material on the web each year.) These services will be recorded only if they become intigral parts of the websites that are included.

One further remark on the task: so far, all websites above federal level and below district level have been excluded from the project. This is especially regrettable in the case of the local associations; the more so as there is a clear tendency for them to publish the elaborately researched chronicles of the association's history directly on the web instead of presenting these as a brochure . Yet, given that about two-thirds of the 12000 local associations present their own website, it seemed an almost undoable exercise under the existing circumstances to include them all. Now, the DFG has submitted a request for us to re-examine whether the local associations could in the future be included in the project. This, however, will cause an enormous increase in the amount of data to be handled. Up to now between 70 and 100 URLs had to be included in the mirroring of the website of the party's Land association, there will now be over 500 (in the case of Bavaria even over 800). At the same time, the volume of data to be archived will grow disproportionately to about 10 times the original amount, reaching 4 gigabytes per Land association. The consequences of this enlargement still have to be discussed with respect to the presentation of the project and the rate of mirroring.
At the moment, we mirror the websites at a two- or three- month interval at national or Land level respectively. In order to avoid losing documents that are put on the web and removed again within the same interval, we are currently developing methods that will allow us to automatically mirror at least each home page in shorter intervals.
On the other hand, at each interval a large amount of data is being mirrored that has already been recorded before. I estimate that at four-month intervals about one-third of the mirrored documents are redundant. However, this cannot be corrected afterwards. The last mirroring of the SPD's websites at national level

comprised about 2 gigabytes. Who can check whether in the meantime there are new links referring to these pages which would become inoperable if the document was to be removed? The notorious sentence: ”collecting is cheaper than selecting, indexing is cheaper than making descriptions and one gigabyte disc space costs a Euro.” (Brewster Kahle) surely does not reflect the entire truth. But deviating from this insight should only be an exception in well-defined cases.
The idea of a continuous mirroring process, which we discussed intensely, seems to me technically impossible to implement – at the moment anyway.

This leads us to the subject of the recording of websites, which we call “mirroring” and which comprises the conventional procedures: acquisition, survey and appraisal. Others may use in the same context the terms: harvest, download or retrieval. No matter which term one decides to use, it always has to signify the physical transfer of an Internet presence into a data structure on a data carrier. The key to this new structure has to be that it remains suitable for browsing by its future users just as we browse the Internet today.

Now, the term ”mirroring“ should not give the impression that this form of recording simply demands a constant entity, for example a server, which will then be mirrored. Such an entity to which we could positively refer exist neither in a physical nor in a logical sense. If it existed, other methods of recording would become possible: as for example the acquisition of entire content management systems or data transfer via FTP. However, as long as the websites are hosted on different servers and as long as not only distinct but also different CMSystems are involved in a single Internet presence, I consider the method of mirroring to be the only viable way of recording websites. In all other cases one would have to re-construct websites on the basis of the acquired contents. This would hardly be feasible or would at least place enormous demands on technology and the time involved. But even though the web does not consist of these single entities, the result of each mirroring process has to be one.

The task that the offline browser, which is our mirroring software, has to perform is to take the chosen section of the Internet and translate it into a complete, fully functional and adequate entity on a data carrier. This demands translating all absolute links into relative links as well as saving all those embedded data files that originate from a section different from the selected one. When talking about the physical transfer of an Internet presence into a data structure, we refer particularly to the translation of links. The offline browser fixes the limit up to which the links will be recorded and the kind of translation from an Internet into a data structure. Hence, interference with the structure of a website becomes inevitable. The rules that regulate these interferences are determined by the configuration of the offline browser. The result is a browser-

enabled copy of the selected Internet section whose authenticity derives from the rules which were followed in the process of its creation.
Of course, there are limits to the mirroring process. Databases, for example, cannot be mirrored, streaming files and session IDs might be problematic. Everything else, however, can be mirrored: dynamically generated sites, java script and even flash animations. But all of this takes place in a constant race between the developers of offline browsers and the web designers. A ready-made solution for the problems related to the mirroring process does not exist – and simply cannot exist.

In addition to the process of recording the form of presentation is crucial because all decisions that have to be made in building up an Internet archive depend on the chosen form of presentation.
After a long period of testing, we found that CD and DVD are less suitable media for presentation. Only the Server Presentation[1] seems to offer an appropriate access to the archive –namely over the archive's Intranet. Only this media assures an adequate reproduction, it integrates the long file names and it can easily be connected to a database like Faust. Therefore, we decided to offer two ways of accessing the Internet archive: one via a small homepage with its own URL and one via the database. (Some Databases already offer respective fields in the input mask with the ability to link-up digital objects and internet addresses)

But the standards of description have yet to be invented. I myself consider every kind of minimalism to be permitted. This is especially because I must advise against the assumption that you'll find metadata[2] in the head of the source code that come close to fulfilling any kind of standard like Dublin core, for example. If there is anything written in the source code, it is so broad and inexpressive – irrespective of the party you are looking at - that it cannot be consulted for description. In any case, in view of the huge amount of data, indexing is the required method of making the Internet archive available. And the description should only complement the index. So we have: The server as media, HTML as datatype, browser as software, and an access via a homepage with an index and/or a database with description.
All the problems mentioned above, such as long file names, index, homepage can be solved for CD and DVD presentation. But this would significantly increase the work load while, in general, providing results of lesser quality.

[1] One has to argue more decidedly in favour of the server as the presentation media, than I did two years ago. See Rudolf Schmitz: Archivierung von Internetseiten / Spiegelungsprojekt im Archiv der sozialen Demokratie(AdsD). In: DA 55 (2002), H.2, S.136 :
[2] Metadata include 1. data created in the process of mirroring (settings, extent, date) 2.a Metatags in the head of the source code 2.b Site information of the remote server and 3. User data. From here on, we only speak of metatags.

The situation is a different one as far as the problem of long-term preservation is concerned. In that case CD and DVD are the appropriate and the most cost-efficient solutions. Because of the long file names, we preserve the mirrored sites in a packed format (Winzip). In addition we back up the data on tapes and use a hard drive with a raid system to duplicate the content.
But the true long-term preservation happens without conversion and without an index because the converted data would lose its functionality. So if you convert the data (in XML for Example) you will need another long-term preservation of that storage format in addition to the presentation format. Whether the additional work and expense we would have to put into this method is justified by an adequate increase in research options, I cannot say today. However, I consider it worthwhile to try and isolate certain formats such as jpg for pictures and certain text forms (press releases) and to preserve them in XML format.
The software we use to archive (Off-Line-Browser, Search-Engine) also has to be preserved as well as the browsers and other tools like Real Player and Acrobat Reader.

The search engine not only creates an unlimited number of indices, it is also able to administer and to combine them. It weighs the results by displaying a result page and does not offer 'dead pages' that cannot be browsed.
Because we mirror in intervals, the combination of different indices assures that you can search through different projects at once. So we do not merge the indices, they stay different, but they are included in one Web form.

When creating the Web form you get to choose between various search options, stemming for example, to search inflections of a word.
(Some of the search options which – for various reasons – we decided not to include as standard are the search for synonyms, the phonetic search and the so-called fuzzy search.)
When displaying the search results all formats except pdf-files will be translated into HTML format and the search terms will be highlighted throughout the document.

Project agenda

I will only name the most serious of the problems that urgently need to be solved and will therefore be the basis of our project work in the near future:

- the automation and dynamic sampling of the mirroring process
The collection of relevant URLs by means of the links to separate pages is the least automated part of the entire archiving process. Tedious and painstaking manual work is characteristic for this part of the mirroring process. This must urgently be replaced by at least partially automated procedures.
- the testing of continuous and alternative recording methods

- determining the possibilities for the recording of ‘deep web’, databases and protected sections of the web like intranets for example and password protected services

Other key areas are:
- the integration of knowledge management systems into the search
- questions about the long term archiving of both the presentation format and the preservation format, as well as dealing with problems of migration caused by to long and anomalous file names.

- and last but not least the development of exemplary criteria of description, data input masks and quotation rules.

Archival value

Whether we are able to prove the suitability of the source category Internet to be archived, will depend on whether we succeed in developing solutions for the problems related to web archiving in the areas of recording, description, preservation and presentation; and whether the technology and time involved in achieving these solutions can be justified. Only the solution of these problems under the aspects of authenticity, research suitability, durability, and user-friendliness open up the possibility of building an Internet archive.
I assume that no one will argue the point that the Internet is worthy of being archived. It is only too obvious that other media are already being marginalised by the Internet.
In any case, political parties are putting more and more emphasis on their Internet presence to communicate with their members and potential voters and to present their agendas and representatives. The possibilities offered by information technology are being systematically included in considerations of both party structure and its working concepts.
In the course of these rapid developments, conventional methods of presentation and communication are increasingly being supplemented or even substituted by Internet services. This is the case at all levels. It applies to a delegate’s letter to his constituency as well as to the organisation chart of a party’s parliamentary group or even to such a central document to the programme debate as the Schroeder-Blair paper, which in fact never was a paper but just an Internet publication.
In direct reference to the Internet, the former SPD secretary general, Franz Müntefering, declares in his article "Demokratie braucht Partei" of April 2000:

“The spread of the Internet as a mass medium will change the conditions of political communication in a radical way within a few years. […] We will strengthen the Internet as the central means of communication within the party.”

"…On and via the Internet all Parties will soon:
-gain, inform and enlist their members;
-manage their members;
-organize independent campaigns suitable to the medium;
-collect the bulk of financial contributions; and
-establish new ways of participation

We want to take an active role in the design of this development, and not only react to it. We will benefit from the Internet by using it to enter into a dialogue with people, in as well as outside the party, to mobilise expert knowledge and to reach those who do not want to work in set structures.
Step by step we will offer innovativ services on the Net, that focus on participation and involvement and that mobilise the resources especially of young members."[3]

Similar statements thave been made by other parties.[4]

The consequential manner in which the parties incorporated the Internet into the strategies of their political work, seems to herald a fundamental change - not only one limited to their communication policies. Decades after the introduction of TV, the political parties in Germany were still unsure whether or how to react to what they sceptically called a 'TV-Democracy'. In the case of the Internet, the parties have at an early stage shown their determination to use this new medium in the spirit of an open and democratic society.

---

[3] URL: http://archiv.spd.de/events/demokratie/muentefering.html
„Die Verbreitung des Internet als Massenmedium verändert jetzt in nur wenigen Jahren die Bedingungen der politischen Kommunikation radikal. [...] Wir werden das Internet als den zentralen Weg der innerparteilichen Kommunikation aufbauen."
"... Parteien werden bald in und mit dem Internet
- ihre Mitglieder gewinnen, informieren und beteiligen,
- ihre Mitglieder verwalten,
- einen eigenständigen, dem Medium gerechten Wahlkampf führen,
- den Großteil Ihrer Spenden einnehmen,
- neue Beteiligungsformen etablieren.
Wir wollen die Entwicklung selbst gestalten und nicht nur reagieren, wir werden die Potentiale des Netzes zum Dialog mit Interessierten, auch jenseits der Partei, zur Mobilisierung von Sachverstand, zur politischen Ansprache derer, die nicht in festen Strukturen arbeiten wollen, produktiv nutzen. ... .
Wir werden Schritt für Schritt eine komplett neue Angebotsstruktur im Netz aufbauen, die auf Beteiligung und Einbeziehung setzt und die Ressourcen mobilisiert, die gerade auch bei jungen Mitgliedern vorhanden sind."

[4] For example by the CDU „Die Entwicklung moderner Kommunikationsmedien und die Möglichkeit, Informationen und Meinungen rasch und preiswert auszutauschen, eröffnen der politischen Arbeit ganz neue Chancen, die es im politischen Wettbewerb zu nutzen gilt. Mit dem öffentlichen Internet-Angebot, dem Mitgliedernetz und dem KandiNet hat sich die CDU diese moderne Entwicklung zu eigen gemacht, die es ständig auszubauen und zu aktualisieren gilt." Und weiter wird von der Notwendigkeit gesprochen, "die neuen Informations- und Kommunikationstechnologien parteiweit zu implantieren"
Beschluss des 13. Parteitages der CDU Deutschlands zur "Reform der Parteiarbeit, 9.-11. April 2000 in Essen
URL: http://www.cdu.de/politik-a-z/beschluesse/reform-der-parteiarbeit.htm