First Monday, Volume 1, Number 2 - 5 August 1996

first monday

PEER-REVIEWED JOURNAL ON THE INTERNET

# THE MYSTERIOUS DISAPPEARANCE OF THE WHITE HOUSE SPEECH ARCHIVE

## A Pioneering Application of Technology Vanishes

By RICHARD WIGGINS

## Abstract

*The White House has been one of the leading government agencies in the United States in using the Internet to publish and distribute information. Staff of the White House have been particularly quick in using new technologies to make different kinds of files available, including audio. For a small portion of this year, an archive of White House speeches and its index disappeared. The files were restored, but their temporary loss raises some important questions over the maintenance of Internet information by governmental agencies.*

## Contents

## A Pioneering Application of Technology Vanishes

Imagine if it were possible to provide American citizens with a full-text archive of the words of our political leaders. Suppose this archive had the following features:

- The full text of speeches would be provided in text form.
- The audio of each speech, in its entirety, would also be available.
- The archive would be indexed in such a way that you could type in any word or phrase you might be curious about, such as "North American Free Trade" or "deficit reduction" or even "Whitewater." The index would allow you to listen to the speech containing the phrase in question, and to follow along while reading a text transcript. You could fast forward through the audio document, or scroll quickly through the text document, as your research needs demanded.
- This archive would be available to all citizens, and others from around the world, via the Internet.

It all sounds fantastic, doesn't it? Maybe by the year 2000 we'll have this sort of thing in place?

Wrong. We already had this sort of sophisticated archive. David Lytel, until recently the Webmaster of the official White House site, conceived and implemented such a service, using some powerful indexing technology and the RealAudio product. And for some reason, sometime in April 1996, the service disappeared from the White House site on the World Wide Web.

## The Internet and Digital Audio

My day job is at Michigan State University, which is the home of the largest academic voice library on the planet, the Vincent Voice Library. We've been experimenting with digital audio transmission on the Internet since 1992. We began with samples of historical voices, including the voices of past Presidents. When the final Presidential debate was held on the MSU campus in 1992, we made a complete audio and text transcript of the debate available on the Internet.

About that time, during a trip to Washington D.C., I had the privilege of meeting Dr. David Lytel, who was working in a policy position at the White House, analyzing National Information Infrastructure issues. I demonstrated some of my university's Internetvoice applications using a Thinkpad laptop in Lytel's townhouse in Washington. I like to believe that this demonstration helped inspire some of the things that came later on the White House site, but of course

Internet-delivered multimedia was destined to explode in 1993 and 1994 without any particular help from me.

Lytel was instrumental in building the White House site on the Web; he describes his role as being the "managing editor" of the site, but in fact he was its principal architect. Thanks to his leadership, the White House was one of the earliest agencies, in any level of government, in any nation, to establish a serious presence on the Web.

Our paths crossed again in 1995, when President Clinton accepted an invitation to give the convocation speech at the university where I work. My colleagues and I set out once again to prepare the 1995 Clinton convocation for multimedia delivery -- text, digital audio, and photographs -- via the Internet. The White House Web site pointed to our archive of this event.

## An Internet-based, Indexed Archive of White House Speeches

A wonderful aspect of digital media is the fact that once you begin to capture something in the digital domain, it becomes easy to build an archive of that material. Another neat aspect is that, with a little bit of effort, it becomes possible to build a table of contents or an index that provides useful random access to listeners. For instance, take a look at the Web version of the 1992 Presidential Debate at Michigan State . If we'd simply digitized the debate and put all 100 minutes of real time into a single file, potential listeners would have to download 50 megabytes of material at once. Such a file would be useless: it'd be too big to download, and it'd contain no index markers to help people find useful starting points.

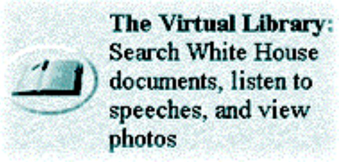
With older technology, one way to solve that problem was to split the larger "document" into many smaller files, with each file breaking at a logical junction. That's exactly what we did with the debate archive: we broke down the 100 minutes of real time into a couple of dozen separate questions, each clearly labeled by topic of the question.

That low-tech solution was surprisingly effective. Suppose we had a contest. One participant is sitting in a room with us in East Lansing; the other contestant is in Miami. Suppose we hand the person in East Lansing an audio cassette labeled "Presidential Debate of 1992." We tell the person in Miami to point her Web browser at the Internet archive of the debate. We offer $1000 to the first person to play back the part of the debate where Ross Perot talked about foreign lobbyists. The Internet user would win easily, even if connected via a dialup modem. That's how powerful an index can be.

Thus, an audio archive is useful, but an *indexed* audio archive is infinitely more useful. The former is a virtual pile of audio tapes; the latter is a library.

Over time, tools for Internet delivery of audio have improved. One tool that caught the attention of the Internet community was RealAudio, announced by Progressive Networks in 1995. Lytel saw RealAudio as a way to build an archive of Presidential speeches, and began by placing Saturday radio addresses online.

RealAudio offers advantages over older formats: Older formats require the user to download an entire file to begin playing it. Even when an event is manually broken up into small segments, this means an entire segment must be downloaded before play begins. By contrast, RealAudio is a *streaming* format; playback begins almost instantaneously after downloading commences. With older formats, even if a speech is broken up into small chunks, it can take several minutes to download data equalling a single minute of real time. With RealAudio, playback can begin within seconds and continue uninterrupted.


The Virtual Library:
Search White House documents, listen to speeches, and view photos

Lytel conceived of a way to combine the streaming capability of RealAudio with a powerful search engine to provide an indexed audio archive. In mid 1995, the White House Web site began constructing a pioneering archive. The White House Web staff licensed search engine technology developed at the University of Massachusetts, and began building an indexed archive of the President's Saturday radio addresses.

The result was something that could be useful for today's citizens and newspaper reporters as well as tomorrow's historians: you could listen to Presidential speeches in their entirety, or you could go back to a particular speech -- or part of a speech -- when you wanted to hear comments by the President on a particular topic. Think of it as C-Span with a brain: unedited speeches delivered in audio format on demand, with an intelligent index, so you don't have to listen through hours of a speech to hone in on a particular topic.

The creation of the Saturday speech archive is an example of how, under Lytel's direction, the White House presence on the Web was much more than the simple "brochure" mode of Web information delivery.

## The Archive Vanishes

I am co-host of a television show about the Internet called *Internet: TCI*. This show is cablecast in mid-Michigan and a number of TCI systems around the country representing almost 1,000,000 homes. During a trip to Washington in March 1996, I taped an interview with Lytel for use on our program.

On Saturday, May 11, 1996, we taped the June episode of our TV show. We wanted to demonstrate how the speech archive works. The only problem is: the speech archive, and the index, had vanished! As we prepared to tape the show on May 11, I could not locate the archive on the White House Web site. I tried conventional browsing of the site; it was not in the "Virtual Library" folder as expected. I tried the "search this Web site"

feature, which core dumped at one point, but it failed to bring up the relevant list of prior speeches and index into same.

THE WHITE HOUSE
Office of the Press Secretary
RADIO ADDRESS BY THE PRESIDENT
TO THE NATION

The White House Web site continued to advertise this facility even after its removal; the label for the Virtual Library continued to say "search documents and listen to speeches," for instance. One could find the previous single Saturday's radio address in audio format, and one could find a few textual transcripts of earlier speeches...but the index of all the Saturday addresses was nowhere to be found.

Being unable to find the archive seemed odd: the archive existed and was populated only a few weeks earlier. As of May 13, 1996, the Frequently Asked Questions document still referred to the archive under its list of improvements to the service; to quote the FAQ:

> *The new service makes extensive use of audio. Not only can citizens search for documents, but they can search the archive of the President's Saturday radio addresses and listen to the section of the speech in which the President talks about what interests them. The new service also includes a database of photos of important events in the Clinton Administration.*

Thus the White House Web site was describing a service that no longer existed on that site.

## So Why Did the Archive Vanish?

So was I missing something? Was the service still there, but they somehow reorganized it and I just couldn't find the pointer? This seems unlikely; I'd looked pretty hard. Lytel left the White House as of April 1, 1996, to explore other pursuits, and he could offer no clue as to why the service would've been removed.

My worst fear was that someone in the White House was afraid that this archive could be used to support "opposition research." That's a process, common in modern campaigning, in which one of your political opponents systematically looks up information from the past that could be damaging to you. One can imagine ways in which an indexed archive of Presidential speeches could be used to quickly locate contradictions (slight or major) in Presidential remarks across time. Just type in a key word or phrase (say, "Bosnia") and you can rapidly locate and review a large number of potentially inconsistent positions uttered across months of speeches.

Of course, in this case, that would be silly. There are numerous existing text indexes of every word the President utters, and you can bet that groups large and small are constantly analyzing his speeches for contradictions. Several months ago I saw an ad on TV that stitched together several sound bites which didn't appear favorable when juxtaposed; they obviously hadn't used the White House Web site to fashion that attack ad.

Although it might be tempting to believe that the Saturday radio archive could be used in opposition research, it really wouldn't be useful in that application: existing text transcripts, available via commercial online indexes for years, would provide the starting points. Furthermore, you'd need broadcast-quality audio and video to build a radio or TV ad, and the RealAudio archive, prepared in a format for delivery to dial-up modem users, would not provide adequate audio quality. In fact, several months ago, when Lytel first told me about his plans for building the speech index, I suggested that the service might be exploited by the President's opponents. Back then, Lytel dismissed the idea as ludicrous; he said the political parties have all the databases and technology they need to accomplish the task on their own. What he failed to realize is that some of his colleagues at the White House harbored that very fear.

## Questions for the White House; the Archive Reappears

On Monday, May 13, I placed some calls to the White House, trying to find answers. I also alerted some members of the press who write about the Internet. On May 13 I also placed online an initial version of this Web page, describing the disappearance of the Archive. As time permitted on May 14 and 15, I continued making calls. Eventually, I reached Rick Borchelt of the Office of Media Affairs, who said that as of April 1 (when Lytel left the government) responsibility for the White House Web site was transferred to the Office of Administration. I was given the name of Frank Reeder in that office, and I left a couple of messages for him.

On Thursday, May 16, I received a voice mail from Mr. Reeder, indicating he'd like to speak with me. I reached him by telephone during the early evening that day, and Mr. Reeder mentioned that he'd seen "the article you wrote." It took me a second to realize that someone must've forwarded the URL for the initial version of this Web page to him; this was the "article" to which he referred. Mr. Reeder indicated to me that the archive had been restored. He said "It was a mistake to remove it. We thought it had the potential for causing problems, but we've thought further and decided to restore it. Thank you for calling us on it. Let us know in the future if we screw up again."

Search Saturday Radio Addresses

Of course, I immediately connected to the White House site.  Sure enough, the radio archive had been restored.

Saturday Radio Addresses

So why did the archive disappear in the first place? I never pressed Mr. Reeder specifically as to whether the reason was fear of opposition research. It seems likely that was the case. Or perhaps someone in the Administration was afraid of more amateurish forms of tinkering -- say, an undergraduate creating bogus sound bites out of bits and pieces of past speeches.

In fact, during Spring of 1996, the Republican party has been running a television ad showing sound/video bites of President Clinton announcing how many years it will take to end the Federal budget deficit. The ad juxtaposes speech fragments where he says "seven years"..."ten years"..."I think we can do it in nine years." Personally, I think the authors of the ad overestimate the power of these minor differences in estimates. In any event, as Lytel surmised, I doubt the White House Web site archive was used to prepare this particular spot, but perhaps it proves the possibilities for such uses.

## Why Government Agencies Shouldn´t Run Their Own Archives

In our interview, Lytel emphasized his pride that the White House site was a *governmental* repository, not a political one; the White House Web presence is pointed to by the Republican Party, the U.S. House, and tens of thousands of other Web sites as a valuable information resource. Lytel recognized his place in history as the first White House webmaster, and he wanted to establish a tradition that the site would house a substantial collection of serious, more-or-less unbiased information.

Unfortunately, the recent experience with the disappearance of the speech archive calls into question the wisdom of having a government agency maintain its own archive of news or historical materials. In this case, we have an Administration that is generally favorably disposed towards electronic dissemination of information: it is said that Clinton has never shied away from a podium -- and this is a sort of virtual podium. Some people credit Vice President Gore with coining the term "information superhighway."

Of course, even this Administration has not always been the friend of full disclosure and free speech. Consider that this President has himself invoked "executive privilege" to avoid disclosure of documents, and he signed without protest the Communications Decency Act.

In fact, while pursuing the question of why the archive disappeared, I was worried what to say if no one at the White House would admit to the *existence* of the archive in the first place. I didn't have a URL for the archive handy, and could only fall back on the testimony of those who had used it. As it turned out, the Administration didn't play games with me; once I found the right person, they not only restored the archive, they admitted that its removal was a mistake in the first place.

But that was this Administration. Imagine if instead the World Wide Web revolution had occurred in 1974, and it was the Nixon Administration that had removed an archive. What are the chances that we'd ever convince Haldeman and company to restore it? (Imagine Rose Mary Woods and the 18.5 minute gap in the digital realm!)

The real question becomes this: can we ever expect a particular government agency to exercise appropriate stewardship over archives of its own information? Although ultimately each agency is expected to work on behalf of the country, not its own leadership, it is naive to believe that agencies do not make choices as to which pieces of information to hoard, which to disseminate, and which to "spin."

"But what if we establish a tradition of dissemination, such as Lytel began?" you might ask. First off, the Web itself is only three years old, and it's hard to make a tradition in so short a time. And in that short time, we've already seen one archive disappear.

Consider another example of a governmental agency that maintains its own archive of speeches -- the U.S. Congress, and the Congressional Record. The time-honored tradition of that document is to allow members the privilege of "revising and extending" their remarks. That innocous phrase allows members to delete passages small and large -- and even to invent entire speeches that were never heard on the floor of the House or Senate!

Traditionally in our society we rely upon libraries to be the keepers of intellectual history. If, in this case, the archive of President Clinton's speeches were under the stewardship of a library -- say, a major research library, or perhaps the National Archives -- potential users of the material would have far less reason to worry that the documents would disappear from the collection.

One myth about libraries is that they keep everything in the collection forever. This, of course, is not possible. They do, however, have "collection management policies" that govern how they winnow a collection. A true research library would never drop a speech archive because it might somehow be used in a way to bring disrepute to someone in power. Research libraries ask two questions when deciding whether to keep something in the collection:

1. "Will this be useful to our patrons?"
2. "Does having this item in our collection contribute to scholarship?"

Note that this story doesn't involve destruction of the only copy of a document collection: the speeches of the President are archived in many places. The question has to do with the *presentation* of the material -- text and audio archive via the

Web -- and the all-important index.

A government agency might decide that a particular set of documents was too sensitive or too injurious to that agency, and therefore conclude that they won't mount the documents on their Web site. Or, an agency might grudgingly provide information online -- perhaps if mandated to do so by law -- but they might do so in a form that is inimical to efficient search and retrieval. By contrast, a library would err on the side of making information available, and would strive to package it in a useful form.

A special group of libraries in the United States has agreed to serve as "depository libraries" -- archives of large collections of government documents. These libraries work together to ensure that citizens have access to *physical* documents; they could also work together to assure broad access to *electronic* documents. One could imagine a plan where significant government sites were "mirrored" periodically at depository libraries. This would allow future scholars to revisit the site *as it existed on a given day.* (Alas, given the realities of server software and CGI scripting, this is not as easy to accomplish as it sounds: each mirror site would have to maintain the same operating system and support software environment as the mirrored site had in place on each day for which a "snapshot" was taken. Still, it's not impossible to imagine.)

So long as the White House maintains its own electronic archive, or the House of Representatives maintains its own archive, or the Department of Energy maintains its own archive... we can expect each organization to behave in its own interest in choosing what to include and what to remove from the collection. This could mean removing significant parts of the collection en masse, or perhaps more selective forms of editing, á la the Congressional Record. One could imagine editing as subtle as a digital incision to remove an expletive muttered by an official under his breath in the course of a speech or debate.

This does not mean that the keepers of the White House site, or of any official government site, are evil people. Frankly, I think it's naive to expect an agency to operate in any manner other than one that puts itself in the best light possible. This includes exercising editorial control over the documents issued by the agency. If my interpretation of how agencies behave is correct, then there is an inherent conflict of interest in having an agency manage its own archives. A trusted third party -- whether it be the National Archives, or a major research library, or even C-SPAN -- has no such conflict of interest.

## Conclusion

The unfolding "Filegate" brouhaha brings to mind another concern about government agencies maintaining their own archives: the privacy of those who search the archives. All Web server software packages maintain detailed log information showing who has visited a site. The "who" information is usually in the form of a domain-style host name (e.g. smith.journalism.princeton.edu) or an IP address (e.g. 35.8.2.23). In either case, it is often possible to tie a host name to a specific individual human with a bit of detective work.

When a government agency mounts a significant Web presence, the keepers of that site could analyze log information in order to see who has looked at which specific documents on that site. For instance, it would've been possible for the keepers of the White House Web site to check to see which pages on their site I'd examined -- and which search terms I'd applied against the site during my searches. Perhaps more importantly, they could have checked to see what *other* members of the media were looking for the speech archive. "Uh-oh! The Mercury News is aware of what's going on. We'd better fix things fast!"

In this case, if such log examination took place, it probably help accomplished my goal: we got the archive restored. But one can imagine other scenarios in which an agency might use logs to build lists of reporters or citizens that seem not to be friendly to their purposes. Such a list could be generated for benign purposes -- perhaps they want to do a better job of communicating with reporters. Or, the purpose could be more sinister: perhaps they want to share such a list with other agencies or perhaps with political operatives with a goal of retribution.

Now let us suppose such an archive were entrusted to a major research library or to the National Archives. I am not perfectly satisfied that all library Web sites understand the sensitivity of their logs. In fact, back in the early days of the Internet Gopher, I discovered a rather shocking breach of privacy at a leading national library -- which was quickly repaired. Nonetheless:

- Professional library associations, such as the American Library Association and the Association of Research Libraries, have strong policies in favor of the rights of privacy concerning "patron" (user) accesses to their collections. In fact, libraries periodically cull their circulation records to ensure that even in the case of a future probe from a government agency -- even for legitimate law enforcement purposes -- only recent, essential records are on hand. The online community and the library community need to ensure that library-based Webmasters understand that these rules apply to their log data.
- Because the motivations of a library are inherently different than the motivations of a political agency, it is exceedingly difficult to imagine a research library compiling any log information to hand over to a government agency. One would predict that a rogue employee engaging in such an act would be fired.

> **Important note:** *I am not claiming or even speculating that the White House staff made any such use of the White House Web site logs in this instance. Instead, I am raising the general question of whether*

> *governmental agencies should maintain their own archives, and, if they do, what privacy guidelines should apply.*

Conclusion

What can we conclude about Dr. Lytel and the White House Web site? We are definitely better off having had Lytel build the service he created. He demonstrated the possibilities for a serious, substantial Web-based collection. The White House site is a model for governmental Web sites. Now we must ask some hard questions: Should we rely on government agencies to maintain electronic archives of important documents? Who will we trust to maintain our digital heritage?

## The Author

**Rich Wiggins**, PO Box 1043, East Lansing MI 48826
517-353-4955 Day Voice / 517-353-9847 fax / email: wiggins@msu.edu

The statements in this page and its descendants represent the personal views of the author. The names of various organizations on these pages appear for identification purposes only; their presence does not imply that they approve of my sentiments. All of these works are copyrighted. Anyone on the World-Wide Web is free to publish hyperlinks to these documents providing proper attribution is given. Any other reproduction or reuse requires the permission of the author.

Based on an original document of 5/13/96
Last major update 5/27/96
Minor additions (including "Filegate" discussion) and grammatical fixes 6/22/96, 6/23/96

## Press Coverage of This Story

POLITICS TRIMS WHITE HOUSE WEB SITE

The San Jose Mercury News ran a short news item by Rory O'Connor on May 15, 1996.

Time Magazine ran a short piece in its Notebook section in its May 27, 1996 issue.
Read the brief Time Magazine article in its entirety.

The Mysterious Disappearance of the White House Speech Archive by Rich Wiggins
*First Monday*, volume 1, number 2 (August 1996),
URL: http://www.firstmonday.org/?journal=fm&page=article&op=view&path[]=472