Archiving Broadcasters' websites

A discussion of web archiving as context to the radio and television collection

Arnoud Goos, MA
Ingest Coordinator
Netherlands Institute for Sound and Vision
Hilversum, The Netherlands
agoos@beeldengeluid.nl

Abstract— **The Netherlands Institute for Sound and Vision decided, after a series of pilots, to begin archiving websites. Its aim is to archive websites that form a complement to its collection, which is mainly focused on audio visual materials and contains hundreds of thousands of hours of Dutch radio and television shows. The Institute believes radio and television is no longer just the broadcasted show. The program is larger and the internet plays a major role in how radio and television are being consumed. In order to recreate an evening of television from 2015 for future researchers, students or other interested parties, one cannot overlook the importance of the internet.**

Index Terms— **web archiving; television; radio; websites**

I. Introduction

The Netherlands Institute for Sound and Vision (NISV) is the national audiovisual archive of the Netherlands, entrusted with the task of collecting, safely storing, contextualizing and making available Dutch broadcasted (public) radio and television programs to media professionals and to the broad public. Its collection contains over 1 million hours of radio, television, film and music, 1.5 million photos and thousands of objects from Dutch radio and television history. The collection is expanding daily through (automatic) ingest of newly broadcasted programs and the acquisition of digital and analogue materials.

After the first internal memos were sent in the Institute about the possibilities of the 'new medium' internet in the mid-nineteen nineties, it still took about ten years before it became clear that internet was nog longer a freestanding medium, but was vastly interwoven with the collection of Sound and Vision. While in the first years of the internet, radio and television makers were using the internet as some sort of expanded TV guide/magazine, later on, the internet was used as an expansion of the program itself. It became a platform where television viewers and radio listeners could access for example the whole interview and not just the broadcasted clip or quote, read a blog of one of the reality soap stars from their favorite show, read background information on the program's subject, and so on.

After a series of pilots Sound and Vision embraced web archiving as a structural part of its collection policy. With this decision, it faced new challenges. For years, the NISV specialized in storing and securing audio visual materials, but how can you archive websites and make sure they can still be accessed decades later? How can you present them to the public? What kind of (copy)rights should be dealt with?

II. Formative years: web archiving pilots

In 2008, the NISV started its first pilot with web archiving. The Institute was a use-case partner in the European project Living Web Archives (LiWa). Sound and Vision wanted to find out what possibilities web archiving had to offer as an addition to its collection. Archived webpages as a way for contextualizing its collection was the main reason for NISV to take part in the initiative.

The three-year LiWa project provided a great basis for testing and getting familiar with the possibilities of web archiving (also called web crawling or harvesting), after which NISV started a pilot project in cooperation with one of the Dutch public service broadcasters. The broadcasting organization, NTR, had to deal with budget cuts and partly as a result of that, television shows were being cancelled and websites had to be taken offline. NTR was afraid the unique content on the webpages would be lost and sought out the national television archive for help. The call from NTR was a good opportunity for Sound and Vision to put the years of research from its R&D department into practice, and also forced the organization to decide what they wanted to do on the subject of web archiving. It made them think about what its role should be in the Netherlands concerning web archiving.

At this time, the internet had already established itself as a complement to the medium of television and radio, as described in the introduction of this paper. Should an archive entrusted with the task of archiving radio and television content ignore the presence of the internet of make an attempt to archive (a tiny part of) it as well? The Netherlands Institute for Sound and Vision went with the latter, which led to a whole new kind of discussion and new difficulties to deal with. Where should the archive draw the line as to what to include in the archive? What is technically possible to archive? And what would or could the institute do with this new part of its collection?

III. Draw the line: the selection of websites

The main question the institute had to answer regarding its selection policy was if NISV should only archive the websites

978-1-5090-0048-7/15/$31.00 ©2015 IEEE

of Dutch broadcasters, or if its horizon should be broadened. In the Netherlands there were already other ongoing initiatives on web archiving, with the National Library in The Hague being the most important player. Audiovisual archives have the responsibility of securing that part of the national or cultural heritage that consists of film, video, audio and still photography. At least that is considered to be the role of the Netherlands Institute for Sound and Vision. But where should Sound and Vision place itself in the national web archiving spectrum? Is a website an audiovisual product? Mostly yes. They combine video, audio and still photography with texts [1].

When the decision was made to include the web archiving project in a larger project called Dutch Weeks of Radio and Television, it helped to shape the form of the project and set boundaries for the selection of websites. During the Dutch Weeks of Radio and Television (two weeks a year, in week number 10, and the week in which the UNESCO World Day for Audiovisual Heritage takes place) Sound and Vision records everything broadcasted on Dutch radio and television. 24 hours a day, seven days in a row, including all commercials and even foreign movies. Realizing it does not have the rights to all the programs, the Institute archives these entire weeks of broadcasting for research purposes only. The content can only be looked within the walls of the institute, making it possible for students and researchers to reconstruct a whole day of television from, for example, the year 1998, covering all stations and networks. Also, as indicated earlier by many different authors on the benefits of web archiving, one can research for example the way language has changed over the course of time, how webpages were linked together, etcetera [2]. With NISV taking up web archiving, the outcomes to these research questions can be linked to their accompanying radio and television shows, to get a broader perspective.

Since TV and radio shows increasingly use the internet for extra content to their programs, it seemed logical to capture webpages from the networks during these Weeks of Dutch Radio and Television. Audiences can watch an extended interview on the broadcaster's website or see the vlog of the show's presenter on what he does on set online, for example. Putting together the list of networks and programs, a question was raised to also include websites that offer a wider look at radio and television culture during those weeks. The decision was made to include a few forums and blogs about radio and/or television, since public debate on the broadcasted shows was mostly held on websites other than the websites of the networks themselves.

When the focus of the web archiving project at Sound and Vision was more or less clear, it became evident the selection guidelines repeatedly need to be revisited. For a medium that is constantly changing, the archive needs to keep track of the changes. A selective collection has certain possible and even inevitable weaknesses. 'The internet does not respect collection and national boundaries', as Maureen Pennock [3] states. Another problem is that of unintentional bias when the selection is made. How to decide what blogs or forums offer a broad view on related discussions online?

IV. LEGAL AND TECHNICAL ISSUES

Like many, or all, web archiving initiatives, The Netherlands Institute for Sound and Vision has to deal with the legal issues that surround the publishing or even the archiving of websites. When the institute started with its current web archiving project, they handled their legal situation just like the British Library did in the beginning of their project. They only archived websites whereof the owners of the website gave their permission to do so [4].

The matter was complicated even more because of legal negotiations between the archive, the Dutch government and the different broadcasters on which role the television and radio archive should have, taking into account the 'new' medium of the internet. Earlier agreements made during the period when the internet was not as open and accessible, would have to be reviewed. Dealing with public broadcasters, the question that is being dealt with is on many occasions, as it is in the case of internet archiving, the following: should the public that paid for the websites through taxes in the first place, not be given full access to what the public broadcasters produce? Why archive TV shows that are made with governmental funding, but not their websites? And should both not be freely accessible? These are questions that are still being debated.

As techniques behind websites develop, a web archiving project should aim to keep up. Well-known technical difficulties in archiving websites are, for example, the use of dynamic content (Flash of Javascripts) and the use of social media feeds embedded on webpages. For an archive that wants to store and secure websites made by broadcasters and capture the public debate surrounding the TV shows they produce, these two issues form the most pressing problems, and deserve attention.

The possibility of archiving social media is one of the ways NISV wants to enrich its web archive. Crawling social media has some technical and legal issues that need to be dealt with though. Many radio and television shows have websites that have an embedded Twitter or Facebook-feed, allowing or encouraging listeners and watchers to discuss the show or comment on it. These discussions are thought to be relevant when trying to archive not only a TV show, but also the cultural and social context in which it was originally aired.

The main problem with these social media feeds is that they often link to another server on which, for example, the Twitter feed is hosted. When the website with the embedded feed is archived, the feeds are rarely archived in their original context. What happens most of the time is that, in the web archive, a code in the subtext links to the current, live Twitter page. For example, Twitter discussions on TV show X on an archived webpage from 2014 can link to the current discussions of the TV show X in 2015. This is definitely not desirable, because it is a form of history falsification.

This technical issue poses a legal problem as well. Does a commenter on Twitter or Facebook know that when they comment on a TV show, their comments will be archived? And that they can possibly be linked to them for all eternity, when they themselves deleted the post the day after because after a

good night's sleep, they felt embarrassed about the post. Of course the legal issue is much larger than this particular problem, but it is one of the many possible examples (for a broader look at web archiving and what the Dutch law has to say about it, see Beunen & Schiphof, 2006 [5]). A solution to this problem would be to only archive the tweets or the posts from the broadcaster itself. But with that approach the possible public debate is not archived at all.

V. SOLUTIONS FOR THE FUTURE

The Netherlands Institute for Sound and Vision acknowledges that web archiving is a continuous, iterative activity. The web archive should keep up with the World Wide Web itself. Websites have developed from standard text pages in the 1990s to increasingly more dynamic pages that are more and more fancier in their appearance. Technical problems should be dealt with and the archive (and the partners they work with) should try to keep up with the developments in website building. The fact that dynamic web content is hard to archive should not be a conclusion, but a starting point for more research to reach solutions.

One of the options the institute is looking into is to try to work together with the webmasters of the websites in its selection. Considering the limited scope of the selection, this would be easier for Sound and Vision than other institutions that crawl many more websites. Besides, a large part of the selection of webpages consists of websites from the public broadcasters they work together with on other levels.

One thing Sound and Vision is working on is to let webmasters know what kind of website features are more archive-friendly than others. For example, it can point out the existence of the Clear method [6] or other tools to show to what degree a website is easy to archive. This does not mean that pointing out the problems will lead to 100% archivable websites for the web archive, but it sure helps.

Working together with the website builders does not only include providing a list with do's and don'ts in programming the website, but perhaps foremost convincing the broadcasters of the importance of a web archive. Broadcasters in general work from show to show and from episode to episode. Yesterday's show is old news. Sound and Vision has to play a part in convincing the broadcasters of the importance of storing and securing the information that gets lost quickly on the constantly moving internet. Otherwise, why would websites be built based on what archives the best, if one does not see the importance of securing the content in the first place?

VI. CONCLUSIONS

The Netherlands Institute for Sound and Vision sees it as its duty to archive not only the radio and television shows that are being aired in the Netherlands, but also to contextualize its collection. Web archiving is one of the relatively new ways to do this, but also, in this day and age, something that can no longer be overlooked.

This means the institute has to deal with issues many other web archive initiatives deal with, but also that it has some challenges of its own due to their unique relationship with the radio and television broadcasters in the Netherlands.

VII. ACKNOWLEDGMENTS

The author wishes to thank Lotte Belice Baltussen from NISV R&D-department for her enthusiastic feedback and helpful comments. Also many thanks to the rest of the web archive team at NISV, as well as the project partners at IMR, Dispectu en Frontwise for providing guidance in for the author uncharted territory.

REFERENCES

[1] A. de Jong, "Preservation of the web: issues for audiovisual archives" Speech at FIAT/IFTA General Assembly Antalya, 13-18th of October 2002.

[2] A. Ball, "Web Archiving (version 1.1)" Edinburgh, UK: Digital Curation Centre, 2010.

[3] M. Pennock, "Web-Archiving", Digital Preservation Coalition Technology Watch Report, 2013.

[4] C. McCarthy, "Digital Libraries: Security and Preservation Considerations" in Handbook of Information Security, Key Concepts, Infrastructure, Standards, and Protocols. Hossein Bidgoli ed. New York: John Wiley & Sons, 2006, pp. 49-76.

[5] A. Beunen, T. Schiphof, "Legal Aspects of web archiving from a Dutch perspective". Report commissioned by the National Library in The Hague.

[6] M. V. Banos, Y. Kim, S. Ross, Y. Manolopoulos, "CLEAR: a credible method to evaluate website archivability" In: International Conference on Preservation of Digital Objects (iPRES 2013), Lisbon, Portugal, 2-6, Sep 2013.