

Strategies and Approaches to Building Thematic Collections in WebArchiv

Libor Coufal
National Library of the Czech Republic
Klementinum 190, Prague 1, the Czech Republic
libor.coufal@nkp.cz

Petr Žabička
Moravian Library
Kounicova 65a, Brno, the Czech Republic
zabak@mzk.cz

Summary

WebArchiv is a joint web-archiving project run by the National Library of the Czech Republic in cooperation with Moravian Library and the Institute of Computer Science of Masaryk University since the early 2000s. First documents were harvested in September 2001. We attempted five large-scale harvests of the whole national domain since then. The archive contains nearly 6TB of data with over 138.5 million documents to date.

The project deploys three strategies of web archiving: automated large-scale crawls of the Czech national web domain; selective harvesting of high-quality websites selected by curators according to a set of criteria; and thematic collections capturing events of national importance. This paper deals with thematic collections only.

Web resources for inclusion in a thematic collection are currently hand-picked by library staff which is a time-consuming and labour-extensive process. When analysing previous thematic collections we noticed that a large part of selected resources comes from a handful of news sites. This paper discusses some possible strategies and approaches for automating the process of building thematic collections including using RSS feeds, alert services and semi-automated harvesting of selected (news) servers. We also present some preliminary findings from a pilot project which tested some of these approaches.

Key words: web archiving, thematic collections, WebArchiv

Introduction

WebArchiv is a joint web-archiving project run by the National Library of the Czech Republic in cooperation with Moravian Library in Brno and the Institute

of Computer Science of Masaryk University in Brno since the early 2000s. First documents were harvested in September 2001. We attempted five large-scale harvests of the whole national .cz domain since then with various, but progressively improving, degree of success. The archive contains nearly 6TB of data with over 138.5 million documents to date.

The project deploys a combination of three different strategies of web archiving: (1) automated large-scale crawls of the Czech national web domain; (2) selective harvesting of high-quality websites selected by curators according to a set of criteria; and (3) thematic collections capturing events of national importance such as elections, natural disasters, cultural events etc. This paper is concerned with thematic collections only.

Depending on the topic, some collections are once-off while others are continuous or long-term by their nature. So far we have either completed or started seven thematic collections, three of them recently (in 2007). Web resources for inclusion in a thematic collection are currently hand-picked by library staff which is a time-consuming and labour-extensive process. It usually starts with using search engines to discover web pages or whole websites related to the topic of the collection. The curator may also try to directly identify websites covering the given topic. The next step involves following links to other related materials. The whole process is then repeated as necessary throughout the duration of the collection.

This cumbersome task obviously raised a question whether it could be made easier. When analysing previous thematic collections we noticed that a large part of selected resources comes from a handful of news sites. By coincidence, around that time we found out that our colleagues from the Danish netarchive.dk project had reached very similar conclusion and came up with a solution that allowed automating the collection process, at least to a degree. However, as it had some potentially serious drawbacks it was not perfect for what we needed. It was an inspiration, nevertheless. We started to look into the matter and identified some possible solutions to the problem:

1. The netarchive.dk approach – regular, repeated automated harvesting of several selected news servers a few levels down from the homepage in short intervals during the duration of the collection; the number of levels along with the duration of the collection need to be determined. The positive of this approach is that it is fully automated; in fact it is a modification of large-scale crawls. As the servers to be collected and the depth of collection (number of levels) can be determined beforehand, collection can be initiated with a minimum delay as the need arises. A big downside is that a large volume of junk is collected during the process, leaving open a question how to provide access to relevant documents only. Also, relevant resources from other servers not included in the collection set are omitted. This drawback can be partially remedied by including topic-specific servers.

2. Cooperation with publishers – all the news servers we regularly use in thematic collections publish RSS feeds that allow us to subscribe to automatically receive new posts from the whole server or its sections. However, these generic feeds are not suitable for our purposes as they still generate a lot of noise and would require curators to sift through the posts to find collection-related content. It would be much better to be able include search terms (keywords) in the feeds that would limit their scope. This approach requires active participation of publishers as specific feeds would have to be created for each collection. The feeds could be created upon a request by publishers based on specifications provided by Web Archiv or directly by WebArchiv and then sent to the publishers. It would slow down the reaction time at the start of collection but the delay in initiating collection can be minimized. The approach would solve the problem of irrelevant content and how to provide access to relevant sources only but it cannot solve the problem of omitting relevant sources from other servers.
3. Using web-based “alerting” services – this approach uses existing free or fee-based web services that allow users to create alerts or RSS feeds for searches. In a way, it is a modification or extension of the previous approach. Rather than having to create individual feeds for each server, it allows users to create only one search feed or alert and use it for repeated searches over a number of servers at once. Some services actually let users specify a set of servers to be covered. Once the searches are created they are periodically run by the service at predefined intervals and the results are delivered either via email or an RSS feed. Examples of such services are Google Alerts or the “Search the future” feature of Bloglines. We will briefly describe and discuss some of these services in the next section.
4. Building an own application – if none of the existing services is suitable it is possible to build an own application. Luckily, it is not necessary to build it whole from a scratch. Many existing services allow users to build “mash ups” by providing application programming interfaces (APIs), for example Google Code or Yahoo Pipes. However, as some programming may be involved it is advisable that the staff working on the project have at least some basic programming skills. Obviously, the more technically apt the staff is, the more complex applications can be built.

A brief description of some web-based “alerting” services

In this section we will briefly introduce two of the services that we call “alerting” – Bloglines and Google Alerts. We have started to experiment with these two services in our thematic collections but we plan to test some other as well in the near future. There is a number of services of this type available that differ in

the features they offer. Many of them are free but some of them operate on commercial basis and charge a fee.

Bloglines

This web-based feed aggregator offers a feature called "Search the future" that lets users to do a search within indexed blogs and then subscribe to an RSS feed that includes these search terms. Anytime new posts appear in the indexed blogs that include these search term, the user gets instantly notified. The feed can be subscribed to in Bloglines or any other RSS aggregator. The search can be further limited to a set of specific subscriptions.

The initial search screen offers only simple "Google style" search box where you can enter keywords with defaulted Boolean AND. However, from the advanced search screen a number of settings can be changed to obtain the best result, including:

- Searching for posts or feeds
- Boolean logic (AND, OR, NOT); phrase search
- Where search terms should appear (in title, author, URL, body, subject, citation, or anywhere)
- Limiting by language and date posted
- Search in all Bloglines or particular only subscriptions
- Results can be sorted by relevance, date or popularity and limited further by time or inclusion/exclusion of news

Google Alerts

Google lets users to subscribe to "email updates of the latest relevant Google results." Alerts are then send to a chosen email account – which doesn't need to be Gmail – whenever there are new Google results for the given search terms. The users can specify whether they want results from News, Web, Blogs, Groups or all of them (Comprehensive). What exactly "relevant Google results" means varies depending on where the results come from. For example, for a 'Web' alert it means "within the top twenty results of a Google Web search", while for a 'News' alert it is "the top ten results of a Google News search." The periodicity of the search can be set to once a day, once a week or continuously. However, alerts are emailed only when new results are discovered. A user can have up to 1,000 alerts at a time. Alerts can be created without the need to sign up but to manage all alerts conveniently from one place it is necessary to have a Google account. The familiar Google search syntax including "phrase search" and all advanced Google search features can be used. The best way to build a complex search query is to use the general Advanced Search page to generate it and then copy and paste it into the Google Alerts search box.

The screenshot shows the Bloglines website's search interface. At the top, there is a navigation bar with links for Feeds, Blog, Clippings, Playlists, Forums, and Search. A search bar is present with a dropdown menu set to 'Search for Posts' and a 'More Options' link. Below the navigation bar is a 'Search Options' panel. This panel includes a 'Search for' section with radio buttons for 'Posts' (selected) and 'Feeds', and a 'Show' dropdown set to '10 Results'. The 'With' section contains four dropdown menus for search criteria: 'All of these words', 'Any of these words', 'None of these words', and 'The exact phrase', each with an 'Anywhere' dropdown. There is an 'Add an entry' button. The 'Language' section has radio buttons for 'English' (selected), 'Any language', and 'A specific set of languages'. The 'Limits' section has radio buttons for 'No Limits' (selected), 'Search only My Feeds', 'Search Bloglines excluding My Feeds', and 'Search the public feeds of this Bloglines user' with an input field. The 'Date Posted' section has radio buttons for 'Anytime' (selected), 'During the last' with a 'Hour' dropdown, 'Sometime' with 'within', 'Any Month', and '2001' dropdowns, and 'Between' with date range dropdowns. The 'Sort by' section has radio buttons for 'Relevance', 'Date, filtered by Feed Popularity' (selected) with a 'No filter' dropdown, and 'Popularity'. The 'News' section has radio buttons for 'Include News' (selected), 'Exclude News', and 'Only News'.

Picture no. 1 – Bloglines “Search the future” advanced search screen

The screenshot shows the Google Alerts (BETA) page. At the top, there is a Google logo and the text 'Google Alerts (BETA)'. A navigation bar contains links for 'FAQ' and 'Sign in'. The main content area is titled 'Welcome to Google Alerts' and includes a brief description of Google Alerts and a list of 'Some handy uses of Google Alerts include:'. Below this, there is a 'Create an alert with the form on the right.' and a link to 'sign in to manage your alerts'. On the right side, there is a 'Create a Google Alert' form. The form has a text input field for 'Search terms:', a dropdown menu for 'Type' set to 'Comprehensive', a dropdown menu for 'How often:' set to 'once a day', and a text input field for 'Your email:'. A 'Create Alert' button is at the bottom of the form. A small note at the bottom of the form states 'Google will not sell or share your email address.' At the bottom of the page, there is a footer with copyright information: '© 2007 Google - Google Home - Google Alerts Help - Terms of Use - Privacy Policy'.

Picture no. 2 – Google Alerts page

Conclusion

Our work on making thematic collections in WebArchiv less labour extensive is just at the beginning. We started to gather ideas for possible solutions and experiment with some of the approaches and services described in this paper. The results we have got so far are only preliminary and they are influenced by the advanced stages of our currently running thematic collections in which not many new resources appear. We will have to wait for a new thematic collection to do a proper test that will give us a better picture. We plan to try out more "alerting" services, as each of them has a slightly different coverage and it can be therefore expected that the best results will be obtained by using a combination of a few of them. We also plan to test the fully-automated approach of netarchive.dk. When we initiate a new collection we will test a few different strategies and compare them against each other as well as with the results of our curators' work. It will be also interesting to see if some of these tools could be integrated with existing web archiving tools such as Web Curator Tool or Netarchive Suite.