Archiving the Web Site of the Bulgarian National Radio - Our Digital Memory Accessible Tomorrow

Anelia Krandeva

Bulgarian National Radio
krandeva@bnr.bg

Abstract. The paper describes some of the issues of web archiving. The preparing the web site of the Bulgarian National Radio for web archiving is discussed.

Keywords: Digital Radio Archive, Multimedia, Radio, Cultural Heritage, Databases, Web Archiving

1 Introduction

Web archiving technology enables the capture, preservation and reproduction of valuable content from the live web in an archival setting, so that it can be independently managed and preserved for future generations. A team from BBC Future Media is in the middle of a project to move millions of pages of BBC Online archive content to a new home. The operation will allow us to switch off an obsolete layer of technology, and secure access to the archive for years to come [1]. It will do integration with multi-platform authoring tools and end-to-end digital production projects.

Semantic Web Archiving System for RF Propagation Measurements and Radio Channel Modeling and Simulation will be built [2]. The system focus is to create an online data repository of wireless propagation measurements. This web service will contain radio channel modeling and simulation tools in various frequency bands, environments, and weather conditions such as indoor, outdoor, urban, rural, broadband, narrowband, mobility, precipitation, etc. The project is currently in early stages of development, but will become an important design and testing tool for wireless researchers in both industry and academia.

Several major projects have been funded for web archiving:

- The SCAPE project (2011 – 2014), developing scalable long-term preservation services for complex heterogeneous and large-scale digital content including web archives. SCAPE runs a Web Content Testbed for testing preservation tools and issues specific to web archives.
- LAWA (Longitudinal Analytics of Web Archives, 2010– 2013), developing tools specifically to analyze heterogeneous Internet data at a massive scale.
- ARCOMEM (2011–2014), focusing on archiving and preservation of the social web, particularly social networks.

- BlogForever (2011–2013), which aims to develop archiving and preservation functionality specifically for content from online blogs.
- LIWA (Living Web Archives, 2008–2011), which explored 'next generation' tools for archiving web content.

Web archives are growing all the time. Whilst some collections remain relatively small, others are now very large – the archives d'Internet at the Bibliothèque nationale de France for example consists of around 300Tb of data, or 17 billion files. Large-scale archives require large-scale search capabilities. This section introduces some of the main tools used for searching and indexing web archives [3].

2 Technical approaches to web archiving

Web archives are born-digital collections that require special software tools for their use and it is a part of the social mission of the Bulgarian National Radio that our web archive collections are preserved and made accessible for the public so for future researchers, historians etc.

HTML was the initial language of the Internet. It has led to the development of several related technologies, including XML (eXtensible Markup Language) and the XML family including XSLT (eXtensible Language Stylesheet Transformations), which is a language for transforming XML documents into other documents and/or formats. Other standards that web crawlers are most likely to encounter include CSS (Cascading Style Sheets), Javascript, and HTTP. HTTP is the protocol for exchanging hypertext and is essential for communicating data across the web.

PANDAS (PANDORA Digital Archiving System) was one of the first available integrated web archiving systems. First implemented by the National Library of Australia (NLA) in 2001, PANDAS is a web application written in Java and Perl that provides a user-friendly interface to manage the web archiving workflow. It supports selection, permissions, scheduling, harvests, quality assurance, archiving, and access.

The Web Curator Tool is an open source workflow tool for managing the selective web archiving process, developed collaboratively by the National Library of New Zealand and the British Library with Oakleigh Consulting. It supports selection, permissions, description, harvests, and quality assurance, with a separate access interface. WCT is written in Java within a flexible architecture and is publicly available for download from SourceForge under an Apache public license.

3 Web site of the Bulgarian National Radio

Established in 1935, the Bulgarian National Radio (BNR) remains the largest broadcaster in Bulgaria. The public radio has two national channels, eight regional stations and ten multilingual channels which broadcast online abroad. For over 80 years, the Bulgarian National Radio has been gathering, recording, broadcasting and preserving Bulgarian history and culture as they happen. The BNR maintains the richest histori-

cal sound archive in Bulgaria, which contains material presenting the oral history of the Bulgarian nation. The Radio Archive is an invaluable collection of sound documents, which can rightly be called as the sound memory of the nation. There are kept indefinitely sound documentation of all significant events of public life in Bulgaria and the world - political, economic, business, cultural, sports and more. The BNR sound archive has a very significant historical value that includes 750 000 sound records, that are equivalent to thousands recorded hours and that were located in thousands of square meters in linear shelves in the BNR Building in Sofia. This archive is part of the Bulgarian and European cultural heritage because it has oral testimonies, as well as broad recorded performances and concerts for the 20th century historical, cultural and society study. This archive is considered the most important audio archive in Bulgaria [4, 5, 6].

The Internet portal of Bulgarian National Radio exists for more than 10 years and is by itself a very rich archive with multiple files, which are an integral part of the Bulgarian cultural heritage.

4 Steps for preparing the Web site of the BNR for web archiving

The main issues are:

- Published content
- Storage and access methods. In December 2001 EBU defined a simple set of metadata which is adapted for use in radio archives, but which is aligned both with the main metadata standards of the broadcasting industry (EBU/SMPTE/AES) and with the Dublin Core metadata (the general approach used by libraries and archives, as well as the worldwide web). Fifteen items of core metadata which are an existing standard (Dublin Core) for Radio Archives are: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights. All the metadata above were taken into consideration when the electronic catalogue of the BNR was created
- Maintaining software/hardware environment. Technical approaches for long-term digital preservation address the challenge of technological obsolescence by ensuring content can be reliably accessed and rendered in the future despite changes to the original hardware/software environment.
- Legal responsibilities
- Selection policies and retention schedules

The problem of the web archiving is very complicate. The radio site archiving has specific problems that we have to be overtaken in the near future.

References

1. Carl Davies, BBC Online@20: Updating the BBC Web Archive, http://www.bbc.co.uk/blogs/internet/entries/f8bd5503-aff3-39fe-9fcb-d88d11e65568
2. Theodore S. Rappaport, Semantic Web Archiving System for RF Propagation Measurements and Radio Channel Modeling and Simulation, http://faculty.poly.edu/~tsr/semanticweb.php
3. Maureen Pennock, Web-Archiving DPC Technology Watch Report 13-01 March 2013
4. Krandeva, Anelia, Emiryan, Silvia, Access to the Sound Archives of the Bulgarian National Radio International Conference on Digital Presentation and Preservation of Cultural and Scientific Heritage, DIPP 2011, pp. 171-177
5. Krandeva, Anelia, Cultural Heritage Archives on Bulgarian National Radio Platforms International Conference on Digital Presentation and Preservation of Cultural and Scientific Heritage DIPP, 2012, pp. 234-240
6. Krandeva , Anelia, Digital Off-Air Radio Events Archive of the Bulgarian National Radio, accepted in Serdica Journal of Computing

Appendix: Web sources for web archiving

- Archive-It: http://www.archive-it.org/
- Archivethe.Net: http://archivethe.net/en/
- DataCite: http://www.doi.org/
- Diigo: http://www.diigo.com/
- HanzoArchives: http://www.hanzoarchives.com/
- HTTrack: http://www.httrack.com/
- Heritrix: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix#Heritrix-Downloads
- International DOI Foundation: http://www.doi.org/
- International Internet Preservation Consortium: http://netpreserve.org/about/index.php
- Internet Memory Foundation & Internet Memory Research: http://internetmemory.org/en/
- Java Web Archive Toolkit: https://sbforge.org/display/JWAT/Overview
- JhoNAS: http://netpreserve.org/projects/jhonas
- MEMENTO: http://www.mementoweb.org/
- Mendeley: http://www.mendeley.com/
- NetarchiveSuite: https://sbforge.org/display/NAS/NetarchiveSuite
- NutchWax: http://sourceforge.net/projects/archive-access/files/nutchwax/
- PANDAS: http://pandora.nla.gov.au/pandas.html
- ReedArchives: http://www.reedarchives.com/
- SiteStory: http://mementoweb.github.com/SiteStory/
- SOLR: http://lucene.apache.org/solr/
- Twittervane: https://github.com/ukwa/twittervane
- Uc3 Web Archiving Service: http://www.cdlib.org/services/uc3/was.html
- WARC Tools Project: http://netpreserve.org/projects/warc-tools-project

- Wayback: http://sourceforge.net/projects/archive-access/files/wayback/
- Web Curator Tool: http://webcurator.sourceforge.net/
- WGet: http://www.gnu.org/software/wget/