The BBC World Service Archive Prototype

Yves Raimond and Tristan Ferne

BBC R&D, London, United Kingdom
firstname.lastname@bbc.co.uk

Abstract. Most broadcasters have accumulated large audio and video archives stretching back over many decades. For example the BBC World Service radio archive includes around 70,000 English-language programmes from over 45 years. This amounts to about three years of continuous audio and around 15TB of data. Unfortunately the metadata around this archive is sparse and sometimes wrong, but on a more positive note the full audio content is available in digital form. We have built a system to automatically annotate programmes within this archive with Linked Data web identifiers. The resulting interlinks are used to bootstrap search and navigation within this archive and expose it to users. Automated data will never be entirely accurate so we built crowdsourcing mechanisms for users to correct and add data. The resulting crowdsourced data is then used to improve search and navigation within the archive, as well as evaluate and improve our algorithms. As a result of this feedback cycle, the interlinks between our archive and the Semantic Web are continuously improving. This unique combination of Semantic Web technologies, automation and crowdsourcing has dramatically reduced the amount of time and effort required to publish this rich archive online. The BBC World Service archive prototype is available online at http://worldservice.prototyping.bbc.co.uk.

1 Description

1.1 Automated interlinking

Between 2005 and 2008 the BBC World Service digitised the contents of its recorded programme library. The digitisation project was a great success but the metadata for it was of limited quality and quantity. It would take a significant amount of time and resource to manually annotate this archive. We therefore consider bootstrapping this annotation process using a suite of automated interlinking tools working from text and from audio.

From text In some cases, textual metadata is available alongside archive content. In the case of the BBC World Service archive, this data could be a synopsis or a title for the programme. In other cases, it could be a script, production notes, etc. We use this data when it is available to try and associate the programme with a number of topics identified by Linked Data URIs.

We process the textual metadata using an instance of Wikipedia Miner [1]. Wikipedia Miner learns from the structure of links between Wikipedia pages and uses the resulting model to provide a service detecting potential Wikipedia links in unstructured text. We trained a Wikipedia Miner instance with a Wikipedia dump from August 2012. Wikipedia Miner returns a set of Wikipedia identifiers for the various topics detected in the text, which we then map to Linked Data identifiers using the DBpedia Lite[1] service. Each of these topics is also associated with a confidence score. We store the resulting weighted associations between programmes and topics in a shared RDF store[2]. For the whole archive, this process generated around 1 million RDF triples, interlinking this archive with DBpedia.

From audio We also use the audio content itself to identify topics for these programmes. This is motivated by the fact that a lot of these programmes will have very little or no associated textual metadata (in the World Service archive 19,000 programmes have no titles and 17,000 have an empty synopsis). And even where textual metadata is present we found it will rarely cover all the topics discussed within the programme.

The full description of this algorithm to extract topics from audio as well as its evaluation is available in [2]. The core algorithm and our evaluation dataset are available on our Github account[3]. We start by identifying the speech parts within the audio content. We then automatically transcribe the speech parts using the open source CMU Sphinx-3 software. The resulting transcripts are very noisy. Most off-the-shelf concept tagging tools perform badly on noisy automated transcripts as they rely on the input text being well written and include useful clues such as punctuation or capitalisation. We therefore designed an alternative concept tagging algorithm which does not assume any particular structure in the input text and can cope with significant noise on the input.

We start by compiling a list of URIs used to tag content across the BBC. These URIs identify people, places, subjects and organisations within DBpedia[4]. We look for possible occurrences of these URIs within our automated transcripts. For example if 'london' was found in the transcripts it could correspond to at least two possible DBpedia URIs: `d:London` and `d:London,_Ontario`. Our algorithm uses the structure of DBpedia itself to disambiguate and rank these candidate terms, and in particular a similarity measure capturing how close two URIs are from each other in the DBpedia graph. For example if the automated transcripts mention 'london', and 'england' a lot, our algorithm will pick `d:London` as the correct disambiguation for the former, as it is very close to one possible disambiguation of the latter, i.e. `d:England`.

Given this technology we estimated that it would take 4 years to transcribe the entire World Service archive on commodity hardware. We therefore devel-

[1] See `http://dbpedialite.org/`.
[2] We use 4store, available at `http://4store.org`.
[3] See `https://github.com/bbcrd`.
[4] See `http://dbpedia.org`.

oped a cloud-based infrastructure to process entire radio archives in a reasonable time. With this infrastructure in place, we processed the 3 years of audio in the archive in two weeks for a pre-determined cost and generated a collection of ranked Linked Data tags for each programme. For the whole archive, the automated audio interlinking generated around 5 million RDF triples, interlinking this archive with DBpedia and the rest of the Linked Data cloud.

1.2 Putting the archive online

We now had an automated set of links for each programme, which we used to bootstrap search and navigation within the archive. The topic data can be used for browsing between programmes, generating topic-based aggregations and searching for programmes on specific topics. We built an application using these links to publish this archive on the web[5]

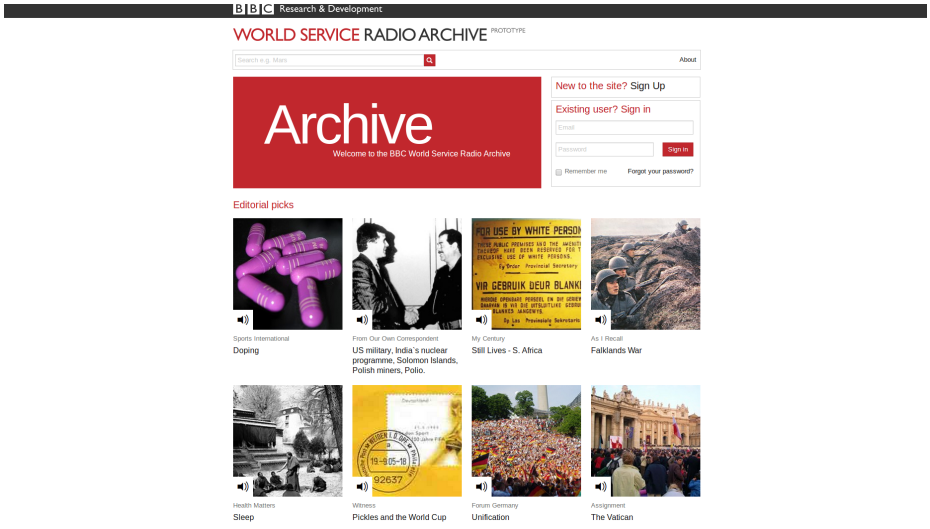

Fig. 1. The homepage of the World Service Archive prototype.

This web site is built using the data held within our shared RDF store. This store includes the automated interlinks mentioned above as well as all the data we could gather around this archive. It also includes a set of images extracted from Ookaboo[6] which we use to illustrate programme pages from the list of topics associated with them. Overall, we store around 20 million RDF triples. Most pages are built from SPARQL queries issued to that store with an average response time of 15ms.

[5] See `http://worldservice.prototyping.bbc.co.uk`.
[6] See `http://ookaboo.com/`

1.3 Crowdsourcing

Automated data will never be entirely accurate so mechanisms are in place for registered users to correct data when it is found to be wrong and to add missing data. When logged in, users can upvote or downvote each individual topic for a programme and add new topics through an auto-completed list, using DBpedia as a target vocabulary. A screenshot of the interface for a 'Discovery' programme on smallpox[7] is available in Figure 2.

TAG	RATINGS	SOURCE
Smallpox	15 0	Synopsis and audio
Infectious disease	12 0	Audio
Eradication	8 0	User
Virus	8 0	Audio
Vaccine	8 1	Audio
Edward Jenner	7 0	User
Public health	7 0	Audio
Disease	6 0	Audio
Infection	6 0	Audio
Weapon	7 1	Audio
Science	5 1	User
Health	4 2	Audio
Laboratory	3 2	Audio

Fig. 2. A set of topics along with their origin and associated user validation data around a 'Discovery' programme on smallpox. Topics can be derived from textual metadata ('synopsis'), audio or can be added by users. When logged in, users can upvote or downvote individual tags by clicking on the thumbs button.

The aggregate of positive and negative votes on each tag is used to improve the machine-generated ranking, and will have an impact on which programmes will be retrieved when a user searches for a particular topic. Gathering this user feedback makes it possible to automatically refine the automated algorithms. This in turns leads to better automated metadata for the rest of the archive creating a useful feedback cycle that leads to a better and better archive experience. As a result of this feedback cycle, the interlinks between our archive and the Semantic Web are continuously improving.

The web site launched in late August 2012 and we are progressively increasing the number of registered users. We now have around 2,300 users. As of August

[7] See `http://worldservice.prototyping.bbc.co.uk/programmes/X0909348`.

2013 these users have validated, invalidated or added over 70,000 individual interlinks. We are currently analysing the quality of the data by comparing tags generated automatically with tags created by expert archivists and general users on a subset of programmes.

As well as refining search and discovery within the archive and helping us improve our algorithm, this user data is also helping us to continuously evaluate our automated interlinking results. The raw user data can be used to evaluate how well our algorithm is doing and we are also tracking the progress of the evaluation measure mentioned above to see how the quality of our interlinks is evolving.

We are also identifying contributors to programmes by segmenting some programmes by speaker, using our diarize-jruby toolkit and an index based on Locality-Sensitive Hashing. We are interlinking these contributors within the archive and with other datasets using a similar combination of automation and crowdsourcing.

We are also providing a visualisation based on an ever-evolving set of interlinks and topics extracted from live BBC News subtitles and described in [3]. This visualisation show archive programmes related to current news events. It enables journalists or editors to quickly locate relevant archive content which can then be used to provide more context around particular events. For example, a recent news event about replacing poppy cultivation by cotton in Afghanistan led to the topics 'Opium poppy', 'Afghanistan' and 'Cotton' being extracted from BBC News subtitles. The visualisation picked up a 2008 radio programme about a new opium ban in Afghanistan and the impact it had on local farmers. An example visualisation is given in Figure 3.

2 Appendix

The application has to be an end-user application The BBC World Service archive prototype is an end-user application, available at `http://worldservice.prototyping.bbc.co.uk`. It gives access to the entire archive of pre-recorded programmes broadcast on the English language part of the BBC World Service since 1947 to any registered user.

The information sources used The BBC World Service archive prototype uses data from multiple heterogeneous sources under diverse ownership and control, including external data sources and automated processes:

- The World Service archive database for the source data and references to audio content;
- DBpedia to describe the topics of the programmes held within the archive;
- Ookaboo to provide visual representations of the programmes held within the archive;
- Topics derived automatically from the audio content;
- Topics derived automatically from associated textual content;

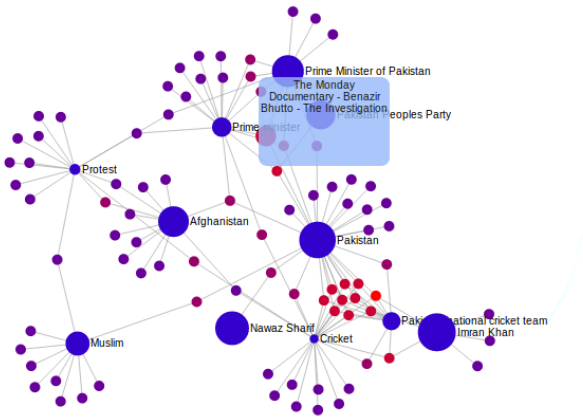

Fig. 3. Visualising archive programmes related to current news events. This capture of the visualisation was taken during the May 2013 Prime Ministerial election in Pakistan (involving Imran Khan, a politician and former cricketer) was discussed on the news. The red programmes in this visualisation include a 1990 Benazir Bhutto documentary and a 2003 Imran Khan interview.

- Automatic speaker segmentations;
- Links between matching speakers across the archive.

The meaning of data has to play a central role Semantic Web technologies play a central role within the prototype. The data backing the prototype is managed as RDF, enabling us to aggregate information from multiple heterogeneous sources, some of them constantly producing new, automatically-derived, data. Moreover, Semantic Web data is central to our interlinking process. We use Linked Data URIs as a target vocabulary for describing the topics of our programmes. This enables us to uniquely identify these topics and to access more information about these topics when needed. We also use the structure of the Linked Data graph as a basis for our automated interlinking algorithm, from audio and from text. We released the core components of this algorithm, generating a vector space from large SKOS hierarchies or large RDF graphs as Open Source. The resulting interlinks are then used to drive search and navigation within the prototype. In particular we build automated aggregation pages around each topic and provide a search functionality. Using Linked Data URIs as topic URIs also mean that other processes implementing the same approach can be indirectly interlinked with our archive, as illustrated by the visualisation described in [3], which displays archive programmes related to current news events.

The application provides an attractive and functional Web interface As crowdsourcing is one of the main aims of the experiment the application was designed for regular BBC radio listeners. The application is available for anyone to register with and has been promoted to mainstream BBC World Service listeners from around the world. Since launching we have had over 2,000 registered users who have listened to around 12,000 of the 36,000 programmes that are listenable and tagged about 7,000 of these. The design process for this prototype has been written about on our blog[8].

The application should be scalable We built a distributed processing framework to be able to apply automated interlinking algorithms to large archives, meaning that the only bottleneck in how quickly we can process archive content is the bandwidth between our archive servers and our processing servers. The Web application uses an index based on ElasticSearch and constructed from the RDF data held in a central triple store to make sure our search engine and aggregation pages are quick enough.

Evaluations demonstrating the benefits of semantic technologies We evaluated the various automated interlinking algorithms, with our datasets and evaluation results available on our Github page. We are still evaluating the impact of user feedback on the average quality of our interlinks.

Applying semantic technology to a new domain or task As far we know, this is the first attempt combining Semantic Web technologies, automated interlinking and user feedback to publish a large archive on the Web. We believe this is innovative in: using Linked Data to generate topics from audio and noisy text data; using machine-generated data and user feedback to improve topic extraction and automated interlinking algorithms; using machine-generated data to kickstart crowdsourcing

Functionality is different from or goes beyond pure information retrieval The prototype offers a search functionality but also offers dynamic aggregations (e.g. on current news topics), visualisations and user feedback functionalities.

The application has clear commercial potential and/or large existing user base The web site launched in late August 2012 and we are progressively increasing the number of registered users. We now have around 2,300 users. As of August 2013 these users have generated over 70,000 individual metadata "edits" (votes, new tags etc). This work intends to demonstrate a cheap and scalable method of putting large media archives online using Semantic Web technology

[8] See `http://www.bbc.co.uk/blogs/researchanddevelopment/2012/11/developing-the-world-service-a-1.shtml`.

and crowdsourcing. There is value in just making archive content available to the public, but also commercial value, for example by re-using/re-purposing archive content for current programming. Being able to quickly locate archive content that may be relevant is key to that, and the overall prototype demonstrates how this can be achieved, by combining automated interlinking and user feedback. This potential for re-use is also the purpose of the visualisation described above.

Contextual information is used for ratings or rankings We use Linked Data available around the topics used to classify our programmes for both disambiguation and ranking within our automated interlinking algorithm. We also use user feedback information to refine these rankings as well as evaluate and improve our algorithm.

Multimedia documents are used in some way The prototype gives access to a large radio archive and combines audio, text and images to describe radio programmes. Some programmes also feature a segmented audio player showing the multiple contributors speaking throughout a programme.

Use of dynamic data Most data within the prototype continuously improves over time, through user feedback and automated processing. A workflow system takes care of the automated processing, and a job processing system takes care of rebuilding search indexes when the underlying data changes.

The results should be as accurate as possible As a result of user feedback and ever improving interlinks between our archive and the Semantic Web, search and discovery within the prototype continuously improve. For example when a user validates that a particular programme is indeed about a topic, it will be more likely for users searching for this topic to find this programme.

There is support for multiple languages and accessibility on a range of devices We are currently investigating opening up this archive to other languages, as BBC World Service broadcasts in many languages. The website is accessible on a wide range of devices, but isnt fully responsive.

References

1. David Milne and Ian H. Witten. Learning to link with wikipedia. In *CIKM proceedings*, 2008.
2. Yves Raimond and Chris Lowis. Automated interlinking of speech radio archives. In *Proceedings of the Linked Data on the Web workshop, World Wide Web conference*, 2012.
3. Yves Raimond, Michael Smethurst, Andrew McParland, and Chris Lowis. Using the past to explain the present: interlinking current affairs with archives via the semantic web. In *Proceedings of the International Semantic Web Conference (ISWC'2013)*, 2013.