Using **machines** and **listeners**

The BBC has an extensive audio archive spanning many decades and is growing at a rapid pace. Making this archive of audio material available to the public poses some unique challenges.

In many cases the metadata associated with each item in the archive is incomplete, inaccurate or unavailable. In addition, current editorial standards or the concerns of rights holders may make publishing certain content difficult. There is also the thorny issue of how listeners can find what they are looking for within an archive where, for example, we may not even know the broadcast date or the title of the programme!

The BBC's approach to opening up archives so far has concentrated around a number of high-profile brands, for example Desert Island Discs, the recently-released Letters from America and themed collections around particular topics. In order to publish these archives, a team of specialist researchers listen to every programme, add tags and other metadata, and then craft a bespoke web-based archive experience to help listeners find what they want. This results in some fascinating and detailed archive experiences, but in each case is a considerably large project to undertake.

When the BBC World Service approached BBC Research and Development with an archive of 70,000 programmes spanning 4 decades, we knew that listening to every piece of audio would be impossible (in fact, it would take nearly two and a half years to do so!). Instead we have used the opportunity to investigate a semi-automated approach to generating better metadata. We analyse the audio to extract as much metadata as possible using speech recognition and audio processing. We then use this metadata, as well as the metadata provided to us by the World Service, to build an archive website. Users of this archive site can then help us to correct and add additional metadata improving the experience for all.

Automated metadata generation

We start by automatically transcribing the audio, using the Sphinx speech-to-text toolkit developed at Carnegie Mellon University. The outputted transcripts have a lot of mistakes. There are many different accents and topics in the World Service, and this can confuse the tool. Take for example this extract from a transcription of a programme:

> *"b. b. c. world service cannot humphrey carpenter presents brittan, comma trade and exploration in the south of countryside and the intrinsic handle the music of the british bid is that benjamin britten s. s. then the sound of the seniors itzhak took the shingle a taste of suffolk one of the east anglia in counties of england at the county to boost the best place to come by ,comma with the company is up Benjamin britten homebuyers hanbury kerr hall handled said seed music with festive for ceding some news from the region and britain's all property ten grimes"*

From this transcript we can discern that the programme may be about the composer Benjamin Britten and refer to some counties within England, but as it stands the transcript is of very little use to understand the full content of the programme. Instead we developed a process to extract some key words, or "tags" from the transcript and use those to interlink and navigate the archive.

One approach that has proved very valuable at the BBC is to use tags that are derived from a controlled, public vocabulary. At the BBC we use Wikipedia terms provided to us by the DBpedia project. If we notice the word "London" in the transcript of the audio, we might tag the episode with the DBpedia tag "http://dbpedia.org/page/London". But what if the audio actually referred to London, Ontario in Canada? In that case we perform a disambiguation step that looks at all the other tags attached to the audio and attempts to chose the best "London". For example if "Westminster" and "United Kingdom" are mentioned in the audio, we can safely assume that "London" in this case refers to the capital of the UK. There are more details about this algorithm on our website [1].

As a result of running this process, the piece of audio is now tagged with a number of concepts which can be searched for or used to link one piece of audio with another. In order to perform this process on all 70,000 programmes in the World Service archive we used a cluster of Amazon EC2 machines. We developed a message-queue based system called Kiwi for this purpose. It distributes the processes across any number of machines and provides an insight into the current status of jobs and easy reporting of any errors. When the worker processes have finished extracting the transcripts from each programme, a final process takes all of the results and aggregates the topic extraction across the entire dataset. It took approximately 2 weeks to perform the transcription and topic extraction on the whole archive.

[1] http://www.bbc.co.uk/blogs/researchanddevelopment/2012/03/automatically-tagging-the-worl.shtml

to publish large **audio archives**

Listeners help to improve the data

We have invited a number of World Service listeners to this prototype (those who are a member of the international panel the World Service uses to test new programmes and ideas called "Global Minds"). After signing in, they are redirected to the homepage of the prototype. This page contains a set of manually curated programmes from the archive, a list of programmes recently listened to, and links to aggregations of topical content in the archive, generated from "on this day" and "in the news" information from Wikipedia. One of the advantages of using Wikipedia as a source of tags is in simplifying the creation of these aggregations.

On individual programmes in the archive, listeners are presented with data coming directly from the World Service archive database if it is present (e.g. synopsis, title, duration, broadcast date), an image, and a set of tags. These tags are derived from the audio, as described above and from analysis of the synopsis of the programme where it is available in our database.

Providing an image for each programme was essential to improve the experience of the prototype, however there were no images associated with each programme in the original database. We turned instead to the free Ookaboo service (http://ookaboo.com/o/pictures/). This service aims to provide a free stock photo associated with every topic in Wikipedia. In our prototype we chose an image corresponding to the most highly weighted tag associated with a programme and then developed an interface that would allow a listener to select a more appropriate image in case the default one was considered unsuitable.

In the current version of the prototype, listeners are able to listen to any programme in the archive using a web-based audio player, and search for programmes using their titles or tags. The search interface is difficult to get right - although we allow the whole archive to be searched, it is important to communicate that the archive itself is incomplete. So a listener may not find a piece of content for a given search term even though the search term is correct, because the content hasn't yet been tagged or had additional metadata added. We provide a "faceted" search to help mitigate this issue, allowing a listener to search within a given date range, or to include or exclude certain tags. We also developed an asynchronous re-indexing process which updates the search index as soon as a new piece of metadata is added. In this way a listener's contributions to the metadata immediately help other listeners to find a piece of content.

To improve the quality of the automatically-generated metadata we allow listeners to modify the tags that have been added to each episode. They can vote up or down tags to indicate how appropriate they are to the episode in question, and add new tags where available in Wikipedia. We are able to monitor the tags that are added and measure how the metadata is improving over time by comparing the tags to those added by a BBC editor. So far in our trial listeners have edited over 6000 tags.

The Future

We are continuing with our research and are exploring some other exciting areas where machines and listeners can work together to improve metadata. We are developing speaker recognition techniques to identify speakers across the archive and then asking listeners to help name them which allows for very powerful aggregations. We are also looking at automated ways to segment audio - looking for sections within programmes that are about specific topics, or identifying music played within programmes for example.

We haven't yet investigated encouraging communities to form around our content and aggregations. We think that by providing mechanisms for listeners to create their own curated collections of content and to apply metadata to those collections we might be able to increase the engagement of the listeners and improve the quality of the metadata as a result. In all cases we are looking to provide listeners easy access to a wide range of exciting World Service archive content but also to provide them with the tools they need to easily correct and add metadata - both to benefit their own experience but also that of other listeners.

We are keen to develop the search interface further in order to reflect the incomplete nature of the metadata more effectively. We are investigating methods to allow listeners to save searches and receive notifications when new content matching their search terms appears. We are hoping that this will result in increased engagement with the archive and a virtuous circle of improvements to the metadata.

Chris Lowis , Yves Raimond
Research Engineers,
BBC Research and Development, U.K.