The Audio Definition Model

Position paper for the Fourth W3C Web and TV Workshop

David Marston
BBC R&D
Centre House
56 Wood Lane
London W12 7SB
UK

The BBC, Audio and the Broadcasting World

The BBC is one of the largest broadcasters in the world, broadcasting and streaming TV and radio over many different platforms. Audio is a critical element of the content, and progress is being made in improving the audience experience in terms of quality, immersiveness and interactivity. As programme production is a world wide activity, where productions are distributed and exchanged, and many different companies and organisations are involved, having standards is vitally important.

Importance of Standardisation

To ensure compatibilty of exchanged programme material we require standards. In the world of broadcasting and production many different standards bodies provide standards and recommendations for audio including the ITU, AES, ISO, ETSI, SMPTE and the EBU. Some standards are unique to a particular standards body, so cause little concern for those using it. Other standards are shared across several bodies and are entirely compatible, and these are also safe to use. The problem occurs when different bodies have different standards for the same things. This is when incompatibilites occur, and they can often happen when exchanging material between different countries who follow different standards bodies. Therefore any new system that requires standardisation should aim to avoid differing versions across different bodies.

The Future of Audio

The multimedia world is moving towards more a involving experience for the audience, with higher-resolution displays, interactivity and immersive audio. For audio there are different approaches to achieving an immersive and interactive experience. Immersiveness (often called 3D audio) is where sound can appear to come from any direction around the listener, including above and below. Interactivity can include the ability for listeners to adjust dialogue levels, change positions of sounds, or select different languages.

To achieve immersive sound, there are three fundamental approaches: channel-based, scene-based and object-based. With channel-based audio, each channel is sent directly to a loudspeaker in a particular location (examples are stereo, 5.1 and 22.2). Scene-based audio represents sound by a combination of dimensional components that combine to make a soundfield, with Ambisonics (and Higher Order Ambisonics - HOA) being the primary technique to perform this. With scene-based audio the soundfield has to be decoded to a chosen speaker layout. Object-based audio represents the sound as separate elements (e.g. singer, drums), and adds positional information to them, so they can be rendered to be played out from the correct location. While each approach has pros and cons, we have to accept that not only will all three be used, but they might be combined in programmes. Interactivity can be achieved by using object-based audio. By sending audio objects separately to be end-user, they can easily control how those objects are rendered. As both channels and scene-based soundfields can be represented by audio objects, it can be considered that all audio can be treated as object-based.

We Need Metadata

The future of audio looks complex, so how do we ensure it can be correctly reproduced for the listener and does not require too much intervention in the production and broadcast/streaming chain. The key is good metadata, and if this is tied closely to the audio then it can allow the audio to be correctly handled and processed throughout the chain. Up until now there lacked any metadata model that sufficiently described the format of audio that is sufficient for these future approaches. Therefore the EBU developed the Audio Definition Model (ADM) (*EBU Tech 3364*) which can give a complete technical description of the audio within a file to allow it to be correctly rendered. So a simple stereo file will have two tracks with descriptions of what Left and Right channels are; whereas a complex object-based audio file with have descriptions for each object so they can be rendered correctly.

The Audio Definition Model

The ADM is a general description model based on XML (but could be extended to other languages). One of its first applications is an extension to the Broadcast Wave Format (BWF) file which includes an 'AXML' chunk to allow XML metadata to be carried. So the ADM is an XML schema which means the audio is described in XML which can be attached to the BWF file. As the ADM is an EBU development it is being added to the EBUCore metadata schema, and will be available in version 1.5.

The ADM is designed to describe the audio as completely as possible, it is not intended to give instructions on how the audio is rendered. For example, a stereo file contains two channels for speakers positioned at -30 and +30 degrees azimuth. The ADM will describe this clearly, what it will not do is tell you what to do if you have a standard stereo speaker arrangement, a wavefield synthesis speaker array or binaural headphones. That is down to the renderer to decide what is best given the description of the audio it receives and the target playout format. However, the metadata should provide enough information for any sort of renderer to acheive its requirements.

The ADM consists of the following elements:

audioTrackFormat	The format of the single track of data in the file
audioStreamFormat	The format of a combination of tracks that need to be combined to decode an audio signal
audioChannelFormat	The format of a single channel of audio
audioBlockFormat	A subdivision in time of audioChannelFormat, allowing dynamic properties
audioPackFormat	A group of channels that belong together (e.g. stereo pair)
audioObject	A group of actual tracks with a given format
audioContent	Information about the audio within a an object
audioProgramme	Information about all the content that form a common programme
audioTrackUID	Identification of individual tracks in an essence

The elements with the Format suffix describe the format of the tracks, streams, channels, blocks and packs in general, but don't describe the audio signal itself. Therefore these definitions can be reused if necessary. For example a file contain 5 stereo pairs (i.e. 10 tracks), will only need two audioChannelFormats ('Left' and 'Right') and one audioPackFormat ('Stereo') to be defined. The other elements cover the description of the actual audio content, so there would be 5 audioObjects and 5 audioContent elements.

So to summarise, the ADM is designed to allow any format of audio to be fully specified so it can be processed or rendered correctly.

Why this matters to W3C

The ADM is in the process of becoming standardised as the audio metedata model for broadcasting, distribution and production. With the convergence of TV and Web technologies it makes sense for the model to become a consideration for the W3C. As the flexibility Web based technology is more likely to allow more varied audio experiences before conventional broadcast technology catches up, the importance of a standardised audio metadata model is something for the group to consider. As one of the topics for this workshop is *'Standardisation of advanced metadata, e.g. for (object) audio or in-band tracks'*, this paper should fit that requirement.