Automatic generation of audio content for open learning resources

Andrew Brasher [1] and Patrick McAndrew [2]

[1] Institute of Educational Technology, The Open University, Milton Keynes MK7 6AA. England
http://iet.open.ac.uk/people/a.j.brasher

[2] Institute of Educational Technology, The Open University, Milton Keynes MK7 6AA. England
http://iet.open.ac.uk/people/p.mcandrew

Abstract: This paper describes how digital talking books (DTBs) with embedded functionality for learners can be generated from content structured according to the OU OpenLearn schema. It includes examples showing how a software transformation developed from open source components can be used to remix OpenLearn content, and discusses issues concerning the generation of synthesised speech for educational purposes. Factors which may affect the quality of a learner's experience with open educational audio resources are identified, and in conclusion plans for testing the effect of these factors are outlined.

Keywords: accessibility, schema, semantic markup, DAISY, audio, digital talking book, usability

Interactive Demonstration: DTBs require a player which will render DTBs conforming to ANSI/ Z39.86-2005, e.g. EasyReader http://www.dolphinuk.co.uk/productdetail.asp?id=9 , or AMIS (adaptive multimedia information system) a free, open source DTB player, http://amis.sourceforge.net/ . For other players see http://www.daisy.org/tools/tools.shtml?Cat=playback,

1 Introduction

E-learning material typically follows an implicit structure, for example by providing items such as introduction, objectives, exercises and support resources. In some cases this structure is also made explicit in the format of the material. Examples of explicitly declared structures include XML schemas such as the XHTML [1] and the OpenLearn XML schema [2]. One benefit of creating and storing material that conforms to an explicitly declared structure is that instances of the material can be transformed to alternative structures using software. In particular for XML structured documents the XSLT language and compatible processors are available [20]. Such transformation is an integral part of OpenLearn reflecting the use of structured documents to author the online materials. Transformations were developed to allow transfer of material in six formats available for download: Moodle, printable HTML, IMS Content Package, zipped collections of resources, IMS Common Cartridge and RSS feeds. One example of how these alternative formats have been exploited is that educators who wish to edit the material, but do not want to engage in editing XML, may now do so by working with one of the other formats (McAndrew & Wilson, 2008). Another example is through the publishing of OpenLearn content in the RSS format, allowing learners to receive courses chunk by chunk (Hirst, 2007). The delivery of the content in the RSS format enables learners to avail themselves of the functionality provided by RSS reader tools. These two examples indicate that using the OpenLearn schema can be of value to both educators and learners.

Our paper describes how value can be added to content by combining the semantics of different schemas to yield instances of learning material which benefit from the semantics declared within the schemas. This is demonstrated by combining a schema developed to describe learning material, with a schema designed to enhance accessibility.

In section 2 a software tool for automating the transformation of simple textual material into a Digital Talking Book (American National Standards Institute, 2005) is described. This transformation combines the teaching and learning semantics provided by the OpenLearn schema with the talking and listening semantics provided by the Digital Talking Book (DTB) schema to yield a talking book with

embedded functionality for learners. For a learner to make use of a DTB a DTB reader is required: a variety of readers and players are available, some being software tools, others being hardware.

Automatic and semi-automatic generation of DTBs can be related to issues within Open Education in at least two ways.

1. Accessibility
   DTBs were originally conceived as a tool to enable those with visual impairments to access printed material; a DTB is a multimedia representation of a resource, combining text and images with synchronised audio. Rendering an educational resource as a DTB can open the resource to those with difficulty accessing visual material, whether it be due to a visual impairment, circumstances (e.g. a learner driving to work), or a learning disability such as dyslexia (Phipps, Sutherland, & Seale, 2002).

2. Sustainability and cost
   It has been suggested that Open Educational Resources should be available at no cost to the user (Downes, 2007). Currently, DTBs are not routinely produced as part of the OpenLearn or OU production processes. Production processes at other universities typically use PDF, HTML or Word documents as the source: these documents usually contain no explicit mark up to identify content which is an important part of a teaching and learning process (e.g. objectives, questions, answers, discussion), and may not contain enough information to allow automatic processes to determine the narrative flow of a document. This means that production of DTBs usually require a significant amount of human resources to hand craft and oversee aspects of the production process. Using content conforming to the OpenLearn XML schema as the input to our process may allow us to automate some or all of the DTB production process.

Consideration of these issues leads to the research questions described in the next section.

1.1 Research questions

The main research question addressed in this paper relates to the accessibility issue:

- What quality factors should be considered when generating Open Educational DTBs in particular and Open Educational audio resources in general?

Subsidiary questions relate to the sustainability and cost issues:

- If we aim for no cost to the user, what investments in production processes need to be made to achieve this sustainably?

- What trade-offs between costs and quality can be made, and how may these affect the learner experience?

The evidence used to answer the questions comes from the development of a prototype transformation tool (section 2), combined with an analysis of literature concerning educational use of audio and textual media. Examples showing how the transformation tool has and could be used to remix open learning content are presented in section 3, quality factors are presented in section 4, and conclusions are given in section 5.

2 The transformation tool: DAISY Pipeline

In this section we look at existing work on the transformation of content into an accessible form and look at how the structured OpenLearn content can be passed into the transformation tool.

2.1 Introduction

The DAISY Pipeline (DAISY consortium, 2007) is open source software for converting various formats into DTBs; DAISY denotes 'Digital Accessible Information System' [3]. The DAISY Pipeline comprises a software framework and API under which a series of individual transformers can be configured, sequenced and run to transform content into a DTB file set, or from a DTB file set into other forms of content. Before describing how customisations of the DAISY Pipeline have been made to generate DTBs for learners, the following two sections briefly describe the output format that is produced by the pipeline (a DTB), and the input that is of interest in this paper (open educational content conforming to the OpenLearn schema [2]).

2.2 Overview of the DTB format

A DTB is a set of files, principally XML files for textual content, control of navigation and text to audio synchronisation (achieved through SMIL [4]), plus audio files in mp3 format. The DTB format is a clear candidate for our application because it improves accessibility, because it can be extended to support specific document types e.g. through use of MathML [5] (American National Standards Institute, 2005; Guillon, Monteiro, Checoury, Archambault, & Burger, 2004), and because it makes use of other standard formats. The ability for it to be extended is important because it means that it has the potential to encode educational resources from different domains (e.g. mathematics, chemistry, literature) without losing important semantic information. In addition, the DTB format's use of other standard formats (e.g. SMIL, MP3, CSS) means that components of the DTB could be reused in other products, or vice-versa.

A schematic of the DTB file set is presented in figure 1. This summarises the main components of a DTB file set, the format of the components, and the relationships between the components. With respect to extensions, if figure 1 were representing a DTB conforming to an extended version of the DTB format the information from the extension would be included in the content files, e.g. the content files could include markup from the MathML namespace.

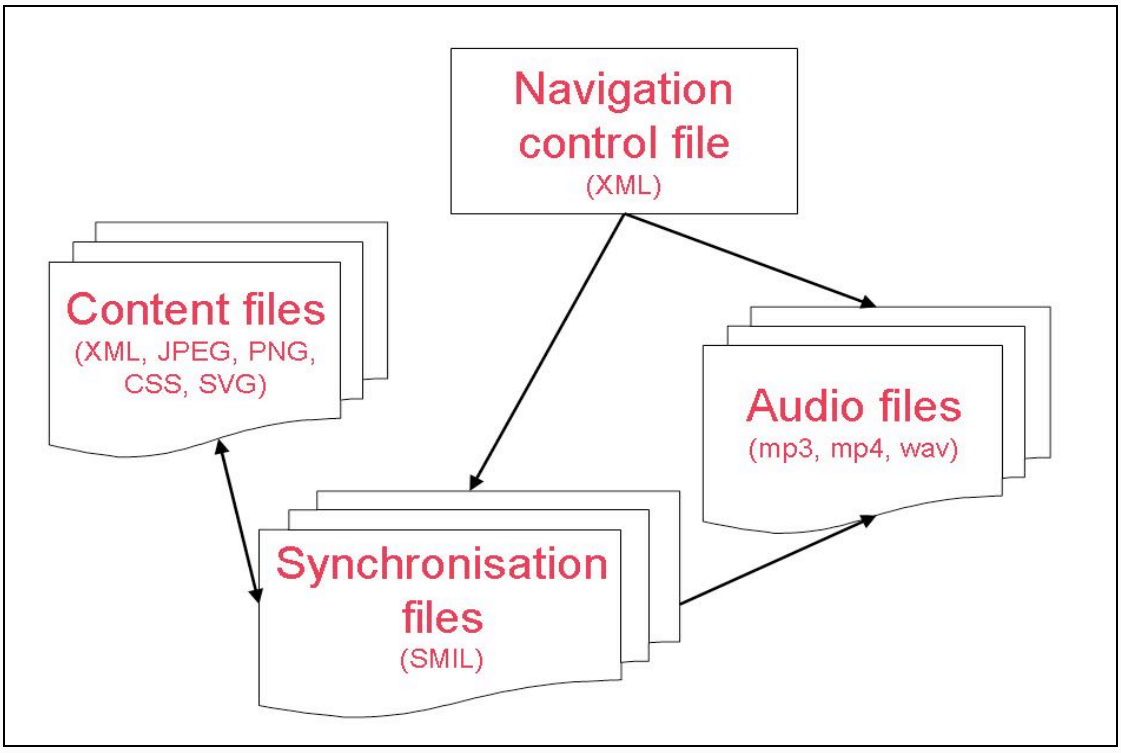


Figure 1: Schematic showing the structure of a DTB

To make full use of the potential functionality of the DTB format requires software or hardware which has been programmed to interpret the format. The navigation control file enables navigation to points within the structure of a book at a macro level (for example, a chapter, or section). The content files can be used to present the content visually to user, enable navigation at a micro level (e.g. between items in a list, or between sentences), and also enable text searching and retrieval operations to be carried out. The synchronisation files enable synchronisation between locations in the content files and locations in corresponding audio files; this enables e.g. a user to search for a phrase and be able to hear the relevant sections of content without having to listen from the beginning of an audio file. Figure 2 shows an example of a content file, SMIL file and audio file from a DTB file set showing how this synchronisation may be achieved: within the SMIL file the <par> element contains multiple elements which should be played back at the same time. The markup vocabulary permissible within a DTB XML content file is similar to that of XHTML, however the DTB content file has the additional "smilref" attribute which enables the relationships between the textual and audio files to be realised as shown in figure 2.

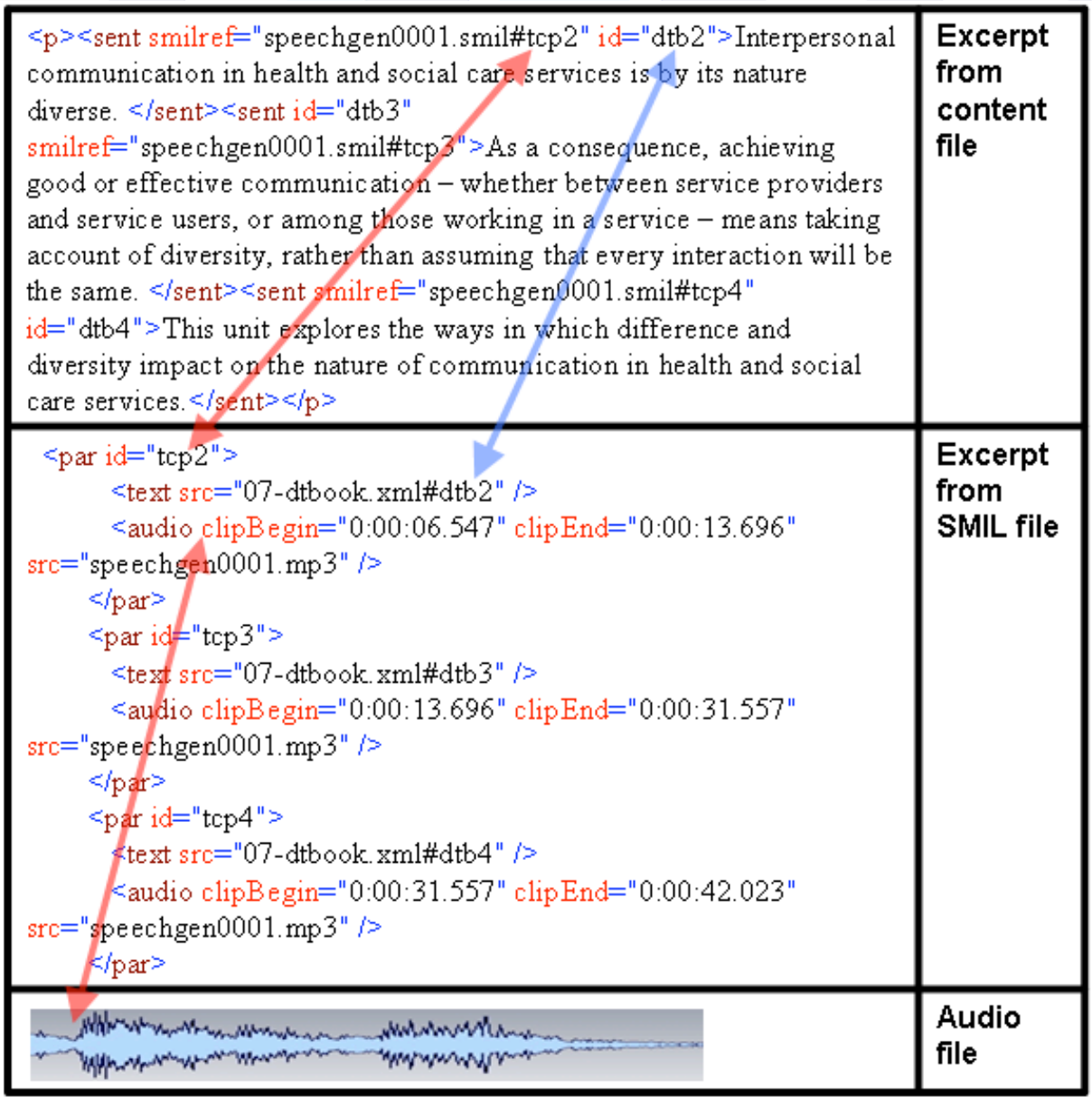


Figure 2: Schematic showing how the relationship between the text in a content file and the corresponding segment of audio file is represented using a SMIL file in a DTB file set

For further information about the format itself see the standard (American National Standards Institute, 2005). For information about how it can be used with players, Leith provides an introduction to how this standard can provide "'a better way to read' for those unable to use standard-sized print" (Leith, 2006).

Figure 3 shows the visual interface of a software reader of the DTB format [6].

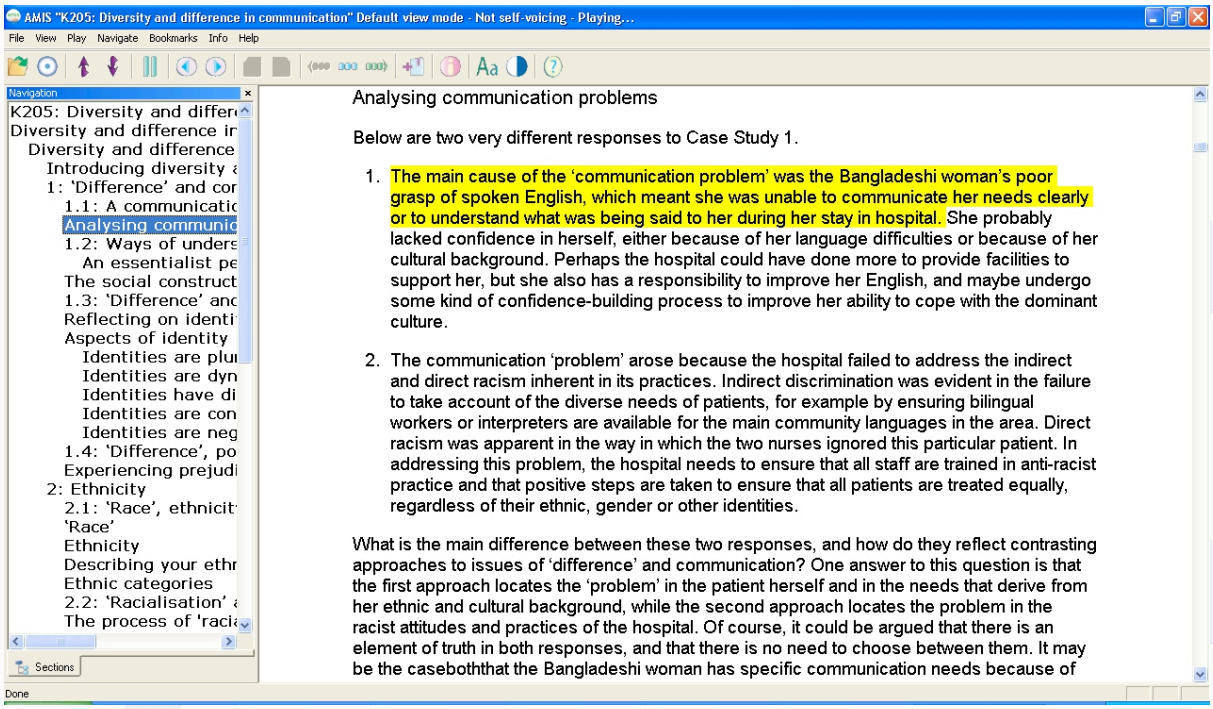


Figure 3: Visual interface provided by the AMIS [6] software

We appreciate that this does not give the full DAISY DTB experience. To experience this, the reader can download an example OpenLearn course unit [7] to try out the features which will be explored in this paper. You will also need software which will play a DTB, for example the open source AMIS software [6], which will play DTBs conforming to an earlier version of the standard (American National Standards Institute, 2002), or there are demo versions of software [13] which will play DTBS conforming to the latest more feature rich version of the DTB standard (American National Standards Institute, 2005).

2.3 Overview of the OpenLearn Structured Authoring schema

This schema [2] provides a structure which enables sections of content to be explicitly marked up as being relevant to a specific educational or learning task. Examples of the mark up vocabulary provided for learning include LearningOutcomes, Activity, Question, Discussion, Example and self-assessment questions (SAQs). This schema has been developed to facilitate reuse of the Open University's course content in multiple media, driven by a business need to standardise and optimise the production of the University's course units. A key aim of the schema was to facilitate the automatic generation of electronic and print versions of course units from the same source files, i.e. from XML source files conforming to the schema. The schema originated from an analysis of the structure of OU course books and has been developed iteratively from 2002 to date (2009).

Intended at first for use only within The Open University, an adjusted version of the schema was publicly released in October 2006 alongside the launch of the OpenLearn site offering Open Educational Resources. In addition to the markup that is related to learning, the OpenLearn schema also provides a variety of tags for marking up the narrative structure within the course unit, e.g. Section, Paragraph, Table, Title and various types of Lists.

An example of a fragment of an OpenLearn course unit marked up according to the schema is shown in figure 4.

```
<Introduction>
  <Title>Introduction</Title>
  <Paragraph>Interpersonal communication in health and social care services is by its nature diverse. As a consequence, achieving good or effective communication – whether between service providers and service users, or among those working in a service – means taking account of diversity, rather than assuming that every interaction will be the same. This unit explores the ways in which difference and diversity impact on the nature of communication in health and social care services.</Paragraph>
</Introduction>
<LearningOutcomes>
  <Paragraph>After studying this unit you should be able to:</Paragraph>
  <LearningOutcome>Demonstrate an understanding of competing perspectives on issues of communication, difference and diversity;</LearningOutcome>
  <LearningOutcome>Demonstrate an understanding of the ways in which issues of ethnicity, gender and disability impact on interpersonal communication in care services;</LearningOutcome>
  <LearningOutcome>Apply ideas about communication and difference to everyday interactions in health and social care contexts;</LearningOutcome>
  <LearningOutcome>Analyse the ways in which ideas about difference can both reflect and reproduce inequalities between groups in the context of care services;</LearningOutcome>
  <LearningOutcome>Identify strategies for working with difference and diversity in the context of challenging discrimination in health and social care contexts.</LearningOutcome>
</LearningOutcomes>
```

Figure 4: Excerpt from an OpenLearn course unit, K205, conforming to the Open University Structured Authoring Generic Schema [2]

2.4 Customisation of the DAISY pipeline

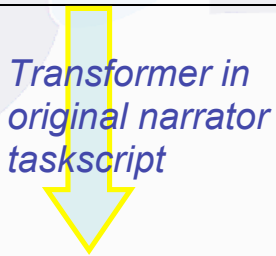
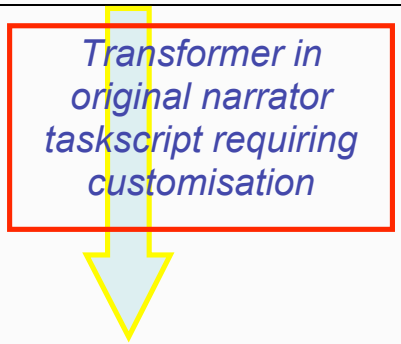
The DAISY Pipeline is open source software for converting various formats into DTBs. The DAISY Pipeline comprises a software framework and API under which a series of individual transformers can be configured, sequenced and run to transform content into a DTB file set, or from a DTB file set into other forms of content. The Pipeline has grammars for describing both individual transformers (the Transformer Description File Grammar [8]) and sequences of transformers, or TaskScripts as they are called (Daisy Pipeline TaskScript Grammar [9]).

A Transformer Description File describes the properties of a transformer, including the format of input it will operate on and the format of the output it will produce. The format of the input that the transformer is designed to operate on and, the format of the output that the transformer will produce are expressed as a MIME types [19]. The Transformer Description File also specifies the Java class that will perform the transformation, and other parameters required for the transformer to operate. The Transformer Description File uses MIME type extensions to specify different XML schemas such as the DTB schema (American National Standards Institute, 2005, [8]).

The taskScript specifies which Transformers are executed and in what order, and it defines the parameters that are sent to the Transformers as they execute. We customised an existing taskScript to operate on OpenLearn Open Educational Resources (OER) by developing new transformers and adding them to the sequence specified by the existing taskScript. The existing taskScript is the Narrator script [10] which creates a DAISY DTB file from a DTBook content file [11].

Figure 5 is a schematic diagram showing the sequencing of the transformers in the customised script. The table below explains the meaning of elements used in figure 5. This script enables automatic transformation of OpenLearn content into a DTB containing features of use to learners. For example, standard DTBs ‘talk’ until they finish or the listener switches them off. The DTBs generated through this script will stop after asking a question within an activity because “producer pauses” (American National Standards Institute, 2005) are introduced at appropriate points by the “Pause inserter” transformer. This means that a DTB player can wait for the user to hit a key e.g. it will not read out the answer or discussion until the learner indicates they are ready by hitting a key.

Table 1 Key for Schematic diagram (figure 5)

Graphic in figure 5	Explanation
	Data operated on by a transformer.
	This is a new transformer we created for transforming OER.
	Data operated on by a transformer, and/or produced by a transformer.
	This is an existing transformer (i.e. one that was part of the original Narrator taskscript sequence).
	This is an existing transformer (i.e. one that was part of the original Narrator taskscript sequence) that will require configuration for most educational applications.

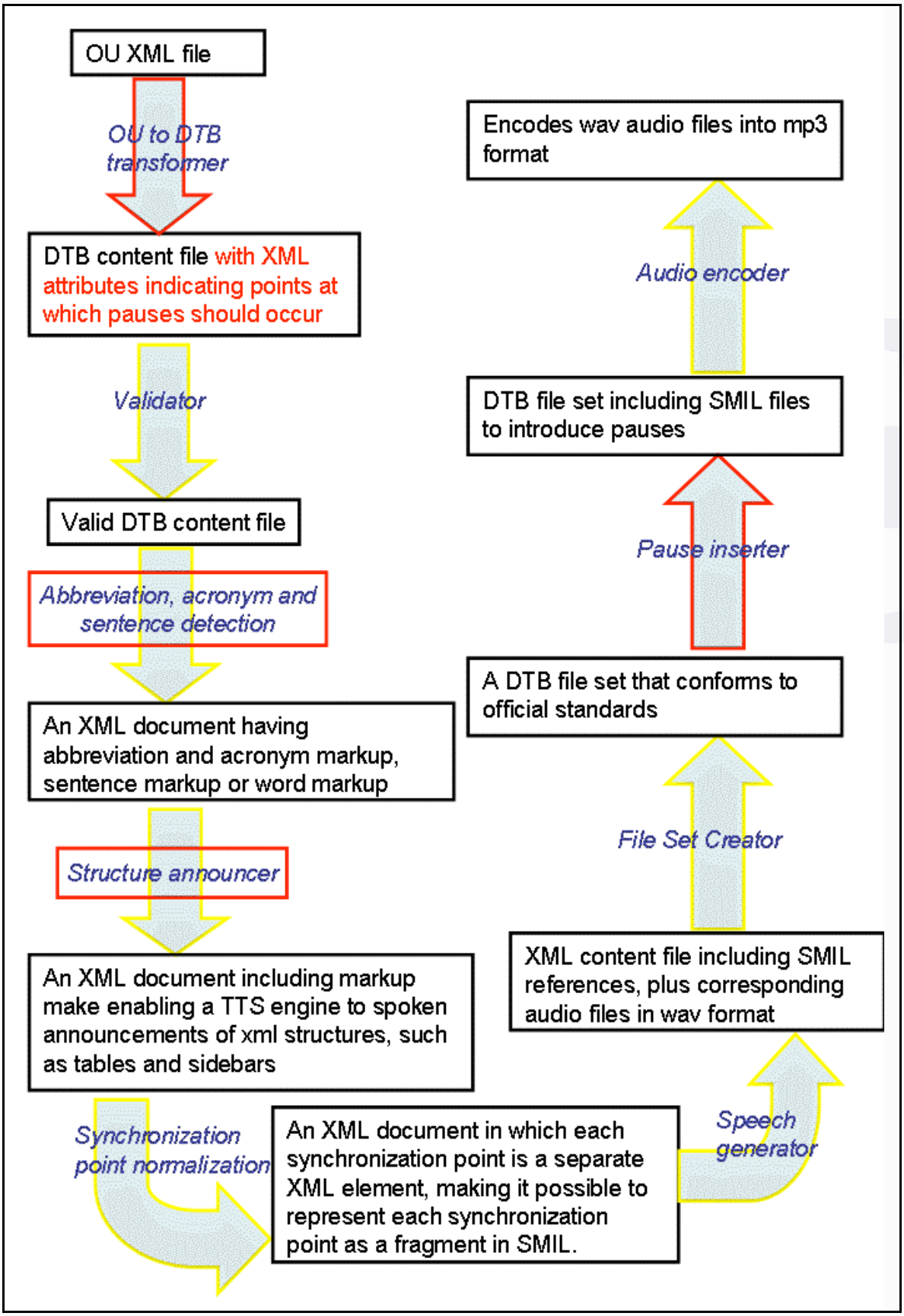


Figure 5 Schematicdiagram showing the sequencing of the transformers in the customised script

Figure 5 shows the positions in the sequence of the two transformers we have added, i.e. the “OU to DTB” and “Pause inserter” transformers. The behaviour of many of the existing transformers can be configured to meet specific needs, and figure 5 also indicates which exiting transformers we believe will require specific configuration for OER by enclosing their names in a red box. For example the “Abbreviation, acronym and sentence detection” transformer can be configured to label text fragments as “initialisms” (for which each letter is pronounced e.g. ‘HTML’), “acronyms” (to be pronounced as a word e.g. “DAISY”), and abbreviations (to be pronounced using replacement text e.g. “for example” instead of “e.g.”).

3 Remixing scenarios

3.1 Introduction

In this section we look at how the extended pipeline can be used in practice by developing three scenarios where content may need to be "remixed" into new versions. We have used Laurillard’s conversational framework to guide the design of two of the remixing scenarios (Laurillard, 2002). The design of the third scenario is informed by evidence from educational use of audio podcasts.

A critical component of the conversational framework is feedback between teacher and learner, because an activity carried out by a learner will be educationally unproductive if the learner does not receive feedback on their performance. Laurillard describes how Open University printed course material often contain features such as SAQs which invite the learner to describe their conception of the topic by writing it down, then to read further, and then to redescribe their conception. This SAQ feature is often used within Open University course material to simulate discussion between a learner and teacher: in sections that the learner should read after they have produced their own answer to the initial question commonly held misconceptions are addressed. In our first remixing scenario we describe how the explicit mark up of SAQs within OpenLearn content enables pauses to be automatically inserted into audio versions of the content at positions which prompt the learner to reflect on the topic.

In the second remixing scenario, we propose educational uses of incorrectly synthesised speech and indicate how the process we envisage could contribute towards sustainably producing audio versions of open educational material. In the third remixing scenario we describe automatic generation of podcasts intended to support revision or initial review of a course unit. This is also an example of automatic restructuring into a format which is widely accessible.

3.2 Prompting reflection using a DTB

In this first remix the start of a question in a SAQ is announced by the utterance of the word “question” in the audio file, and the end of a question by the utterance of “end of question”. Provided that the software playing the DTB includes the functionality to respond to producer pauses, playback will stop after the question; playback will be resumed when the user hits a key (or some other user interface mechanism afforded by the playback software or hardware). This pause invites the learner to take their time and reflect, before the content of the <Discussion> tag is played. Once they have listened to content of the <Discussion> tag the learner can revise their conception of the topic in view of the points made by the ‘teacher’. If the software or hardware does not include functionality to respond to producer controlled pauses, playback will continue without interruption.

An example of markup in the OU OpenLearn resource is shown in figure 6 illustrating the markup that is used to introduce features of use to learners. The course unit we have used to demonstrate this remixing scenario is “Diversity and difference in communication” [12], an introductory level course which is part of the OU’s OpenLearn offering under the theme “Health and lifestyle”.

Markup excerpt from OpenLearn XML for K205 course unit	*Notes*
</CaseStudy> <Activity id="ACT006_001"> <Heading>Activity 1: What is the problem?</Heading> <Timing id="TIM006_001"> <Hours>0</Hours> <Minutes>20</Minutes> </Timing> <Question> <Paragraph>The speaker in case study 1 above is a Bangladeshi woman living in the UK. Having read the case study, think about the following questions.</Paragraph> <NumberedList> <ListItem> <Paragraph>What is the nature of the communication 'problem' experienced by the speaker?</Paragraph> </ListItem> <ListItem> <Paragraph>Whose problem is it?</Paragraph> </ListItem> <ListItem> <Paragraph>What are the consequences for the speaker?</Paragraph> </ListItem> </NumberedList> </Question> <Discussion> <SubHeading>Comment</SubHeading> <Paragraph>Note that the speaker herself identifies the existence of a 'problem' and relates it specifically to 'communication'. Initially, she analyses this as a problem of 'language'. Presumably	*The opening tag of the <Question> element is used by the Pipeline to introduce the word "Question" into the DTB audio, but not the text.* *The closing tag of the </Question> element is used by the Pipeline to introduce the phrase "end of question" into the DTB audio*

(although we do not have all the facts) she is referring to her perception that her spoken English is not sufficiently proficient to enable her to express her feelings as she would like to. She goes on to mention a specific experience, in which she was in hospital for two weeks. Here the precise nature of the problem becomes rather less clear. She says that two of the nurses 'neglected' her, and that she is unsure whether this was 'because of my colour' or because of the 'communication problem', by which presumably she means the language issue she mentioned earlier.</Paragraph> <Paragraph>This raises the question of whether the real communication 'problem' was the woman's inability to speak English, or the racism of some of the hospital staff, which resulted in them failing to communicate important information to her. That racism might have been expressed in a very direct way – simply by ignoring the patient 'because of my colour' – or more indirectly, by failing to provide the services (perhaps a bilingual worker or an interpreter) that would have made communication possible.</Paragraph> <Paragraph>Whatever the nature of the 'problem', the consequences were potentially very serious, in that the speaker was discharged without knowing the precise nature of the surgical procedure she had undergone. A failure or breakdown in communication can lead directly to poor quality care being provided.</Paragraph> </Discussion> </Activity> <Paragraph>This brief case study demonstrates both the importance and the complexities involved in issues of communication and difference in the context of care. The questions were not straightforward, and they were designed to show the complex and contentious nature of the issues, rather than to produce easy answers. In a sense, how you answered the questions in Activity 1 depends on your understanding of the nature of 'difference' – whether of ethnicity, gender or disability – and both how difference is produced and how it should be responded to.</Paragraph>	*(but not the text). It is also used to position a "Producer controlled pause".*

Figure 6: Example of markup of an activity with an OU OER showing which tags are used to introduce features of benefit to learners into a DTB

One issue that we have not discussed so far is the inclusion of material which is external to the OU OER XML but that is part of the unit of learning, e.g. PDF or other files. For example, the PDF file K205_1%20Reader%20Chap%2012.pdf which is part of the "Diversity and difference in communication" course unit [14]. In principle there is no reason why external resources such as these could not be converted into DTBs and automatically linked to the main course unit DTB. The ease of such a conversion would depend on the format of the resource given that the DAISY Pipeline already provides facilities for automatic generation of DTBs from formats such as XHTML and Word XML.

3.3 Working towards an adaptive remix

Laurillard classifies computer-based media which change their state in response to a user's actions as 'adaptive media'. She describes how media which conform to this classification provide intrinsic feedback to the learner i.e. feedback which is internal to the action the learner takes, and cannot be stopped once the action occurs. This contrasts with extrinsic feedback which is external to the learner's action and typically occurs as a teacher's comment on the action. To be useful to a learner any intrinsic feedback provided by an instance of adaptive media must be meaningful to the learner, in that the learner must find it easy to interpret in relation to the goal they are trying to achieve (Laurillard, 2002). Laurillard states that an adaptive program is

> "one that uses the modelling capability of computer programs to accept input from the user, transform the state of the model, and display the resulting output".

We propose that a speech synthesizer could be used as an adaptive component to support learning using a DTB. In this proposal, the model is the speech synthesizer's rendering of the text and the resulting output is 'displayed' as sound. To be an adaptive medium, a speech synthesizer would have to be controllable in some way by learners. In our proposal we hypothesise how an online speech synthesizer could be used.

Learning through proposing phrases to fit context - a remix which omits particular words or phrases

Consider the OpenLearn unit "Spanish: Con mis propias manos" (L314_1). As this is a level 3 course, most of the text is in the language to be learnt i.e. Spanish, and the learner's expected first language English is used only occasionally. The remix begins with a remixer picking particular components of the language which are used within the unit. These components could be individual words or phrases, or parts of verbs. These components are supplied as input to the remix engine, which then generates a DTB in which all instances of the selected language components are replaced by 'x's in the DTB text and by silences in the DTB audio. The silences generated will be exactly the same duration as the synthesised 'spoken' component would have been.

This remix would be made available to learners with instructions that their task is to listen out for silences, and to suggest appropriate words or phrases to fill the silences. A typical interaction might be that a learner listens to the remix and, as they are listening, writes down a word which they think is suitable to fill every gap they perceive. Once they have completed that exercise, the learner visits a URL suggested to them in the initial instructions in the DTB. At this URL they find a form into which they can enter every word they have noted down (or e.g. the figure zero to indicate that they missed a particular word). They then push a button which causes a new version of the passage to be synthesised with their words in the DTB text in place of the 'x's, and spoken in the synthesised audio in place of the silences. The learner can then listen to the audio from this newly generated DTB and reflect on whether their choice of words and/or phrases is appropriate. If not, they can revisit the form, edit the set of words/phrases they supplied and then generate another version of the DTB. A learner having difficulty identifying a particular missing word would be able to listen to the word in its original context in the complete version of the DTB's audio.

An obvious application area for this idea is foreign language learning, though it is relevant for any domain for which interpretation of new vocabulary in spoken communication is necessary.

3.4 Podcasts for revision

The design of this remix has been informed by two studies of podcasting in education, and an analysis of the effect of the rate of synthesised speech on comprehension.

Evans (2008) describes a study of undergraduate students' use of podcasts for revision. In this study, each podcast consisted of a 5 minute long MP3 audio recording of the course lecturer reviewing the learning outcomes and adding clarifications. Three podcasts were produced for first level students of an ICT course, and these podcasts were made available to the students at one week intervals after teaching had finished but before the end of course examination. Evans' results focus on students' perceptions of podcasts, and they suggest that students find podcasts to be efficient, engaging and easy to use for revision.

In their paper on podcasting for learning, Cebeci and Tekdal (2006) recommend that educational podcasts should be less than 15 minutes long, based on a survey of students' opinions about use of MP3 players in education.

Reynolds and Givens (2001) studied the effects of rate of presentation on comprehension for both synthesised and natural speech. Their results suggest that rate of presentation (i.e. speed of speech) within the natural average range of 135 up to 187 words per minute should not have a significant influence on novice users' ability to process and comprehend synthesised speech. Reynolds and Givens recommend that, when choosing the presentation rate of synthesised speech for an application, the preference of the individual user is the most important factor.

With this in mind, we have designed a version of an OpenLearn unit which can be automatically realised as a DTB; the aim of this particular version is to give learners who have interacted with a particular unit relevant prompts to aid them in refreshing their memories about the concepts within the unit.

By combing these findings from the three papers, we propose a structure for a revision podcast that can be automatically generated from OpenLearn course units. Podcast formats such as RSS and Atom enable chunks of audio to be combined into a feed to which a learner can subscribe, and the learner receive each audio item in sequence, as the are published. Although the syntax of the variants of RSS and Atom formats differ, the semantics are broadly similar (Wittenbrink, 2005). They all offer the ability to deliver channels of information, in which each channel can contain a title, description, publication date, and any number of items. Each item can contain a title, a description, publication date and a link to an audio file. Our structure for revision podcasts is shown in figure 8.

```
<?xml version="1.0" encoding="UTF-8"?>
<rss>
        <channel><title>Revision podcast for the unit Diversity and difference in
communication</title>
        <link></link>
        <description></description>
                <item>
                        <title>Introduction and Learning outcomes</title>
                        <description>This will contain text from the introduction, and all the
learning outcomes</description>
                        <enclosure url="./speechgen0001.mp3" length="97"
type="audio/mpeg"></enclosure>
                </item>
                <item>
                        <title>Activity 1: What is the problem?</title>
                        <description>Text from first activity</description>
                        <enclosure url="" length="" type=""></enclosure>
                </item>
                        <item>
```

```
                    <title>Discussion of activity 1</title>
                    <description></description>
                    <enclosure url="" length="" type=""></enclosure>
                </item>
        </channel>
</rss>
```

Figure 8: XML code sample showing proposed structure of a revision podcast

The number of activity question/discussion item pairs would be limited so that the length of the podcast is no more than 15 minutes at a typically normal speech speed i.e. 160 words per minute. This would be done by the generation algorithm, as the duration of each item would be available to the algorithm after the speech has been synthesized. Alternative remixes could be produced at different speech rates allowing learners to choose a rate that suits them. Remixes encoded at higher rates (i.e. greater than 160 words per minute) will potentially contain more question/discussion item pairs than those synthesized at lower rates.

4 Quality, sustainability and cost factors

Carrying out the work described in sections 2 and 3 has enabled us to put forward factors which we believe will affect the quality of a learner's experience with automatically generated open audio remixes. An overview of the quality factors we have come up with is presented in section 4.1, followed in section 4.2 by an overview of factors related to the sustainability and cost of generating audio remixes. To date we have not carried out any usability tests or other investigations with learners to determine the relative impact of these factors, but our plans in this respect are outlined in section 5.

4.1 Quality factors

Expressiveness of the schemas

Are the schemas expressive enough to encode the information required for their users' needs? For the OpenLearn schema the 'user' will be a teacher, and for the DTB schema the user will be a learner. In other words,
(i) does the OpenLearn schema enable a teacher to accurately express their pedagogic intent, and
(ii) does the vocabulary of the DTB schema enable that intent to be expressed so that players will enable a learner to experience it via audio?

Audio content factors: language, voice and speech style, synthesis

The content of most OpenLearn course units has been designed and produced with the intention of delivering it visually, either to print or via the web. The grammatical style and language used in the text have been chosen with these media in mind. Is taking the content and in effect reading it out the best that can be done? One example of a way in which the content can be changed for audio presentation was described in section 3.2: in our prototype our algorithm inserts prompts that do not exist in the original text e.g. "question", "end of question". How useful are these to a learner? What other sorts of changes should be automatically made to benefit the learner?

Rowntree (Rowntree, 1994) discussed the language, style and voice to be used for audio teaching material. He suggests that using different voices (e.g. .a man and a woman's) may keep the listeners' interest. This prompts us into considering how the speech synthesizer(s) within the pipeline should be configured. Though the quality of the voicing produced by modern speech synthesizers can be very good, the nature of some higher education material can mean that customisation of the synthesiser is required. For example, some disciplines make use of specific vocabularies, acronyms and pronunciations. This means that although the fully automatic process will often yield audio content

which is suitable as a prototype, manual intervention to configure the voice synthesizer for different disciplines may be required to achieve high quality 'readings' and to give helpful prosodic cues to the listener. If using different voices as suggested by Rowntree is found to be beneficial, a switch in voice could be done automatically by the pipeline. For example, quotations could be delivered in a different voice from the main text.

Other techniques could be beneficial for particular kinds of content, including the application of auditory display methods (auditory display has been defined as 'the use of non-verbal sounds to convey information' (Kramer et al. 1999)). For example, Brown et al. suggest various ways to present tables (Brown, Brewster, Ramloll, Burton, & Riedel, 2003), and Barrass has identified several design patterns from work done by auditory display specialists, e.g. the PerceivingPatternsInData pattern which is intended "to support exploration and discovery of patterns in complex multi-attribute, multi-dimensional and/or time-varying data" (Barrass, 2003; Stephen Barrass and other contributors, 2003).

Player features

The capability and usability of the device used to deliver a resource utilising audio to a learner will obviously affect the quality of the learner's experience. It should be remembered that a DTB combines visual content with audio content; Nes notes in her study of the use of DAISY for print disabled students in Norwegian Primary and Secondary education, that a well-designed graphical user interface will benefit the non-visually print-disabled (Nes, 2007). Podcasts can also contain structured text and images encoded using XML (e.g. RSS, ATOM [17]) that can de displayed by a player.

Diagrams, images and other visual material

Many OER (including OpenLearn course units) contain digital images or videos. Audio descriptions of images and videos can be automatically generated, if textual descriptions are available e.g. within the OER. Taylor provides guidelines in use at The Open University for describing visual teaching material for students learning at a distance (Taylor, 2004), and the descriptions of images and videos in OpenLearn units will most likely have been produced in accordance with these guidelines. If human produced textual descriptions are not available, it is unlikely that they can be automatically produced for every image or video. However, it is possible that auditory display techniques can be applied in certain circumstances. For example, if the source data used to produce an image showing a graph is available, one of the auditory graphing techniques described by Flowers could be applied (Flowers, 2005).

4.2 Cost and sustainability factors

In their analysis of sustainability of open education Atkins et al argue that one factor that can contribute to long term sustainability is the generation of open educational resources as a low marginal cost derivative of routinely used course preparation systems (Atkins, Brown, & Hammond, 2007). The (semi-)automatic generation of DTBs and other audio formats from OU XML course units could fit this model, because course units in OpenLearn XML are routinely generated by the OU's production systems. There are several factors which will affect how low the cost the audio derivative can be. These factors include

- **Compatibility and evolution of schemas**
  As the schemas evolve, further implementation costs will occur to maintain and improve the ouputs that can be generated through automated production techniques. Evidence that the schemas have and will change is provided by the announcement in January 2008 of "an upgraded version of the OpenLearn XML schema"(OpenLearn, 2008), and the requirements gathering process for the next version of the DAISY standard which runs from October 2007 until March 31, 2008 (DAISY Consortium, 2008). Different schemas can express the same semantic intention using different markup. When this type of variation occurs, a transformation from one markup to the others needs to be implemented by a programmer. One example of this is the markup for tables in the OpenLearn and DTB schemas.

- **Popularity of chosen implementation**
  Sustainability may be influenced by the number of learners who can access a particular format of remixed OER containing audio, and the ease with which they can do so. This will be influenced by the popularity of the format(s) used to deliver the resource. For example, podcasts utilising RSS/MP3 may be more popular because there are more installations of the software needed to play them.

- **Synthesizer configuration costs**
  Modern text-to-speech synthesizers use a combination of methods to pronounce different words within a given text. Typically pronunciations of frequently used words are looked up in the synthesizer's dictionary. For less common words which are not in the dictionary, e.g. names of people or products, there are rule-based and data driven approaches to automatically generating the pronunciation. These automatic approaches used to generate the pronunciations of other words may or may not pronounce every word as expected. Problems with automatic pronunciation are described by Bellegarda (2005), Damper, Marchand, Adamson, & Gustafson (1999) and Spiegel (2002). In their studies of a variety of automatic methods Bellegarda and Damper et al. both report the lowest error rate in automatic pronunciation of words as being around 30%. In other words, a typical speech synthesizer will pronounce correctly only about 70% of words that it does not already 'know', though this figure varies with the language being synthesised because of differing complexities in spelling-to-sound correspondence. To ensure that the audio rendered by a speech synthesizer is error free requires human intervention to check the audio, and if necessary to edit the synthesizer's dictionary or to program the synthesizer's pronunciation generation algorithms.
  Techniques such as use of learner input to identify mispronunciations could also contribute, by reducing the need for expert intervention. This could be achieved by asking skilled learners to identify mispronunciations in synthesised audio before it has been checked by experts, with the aim of reducing the experts' workload. The benefit to the learners who participate would be feedback from the experts.
  Free exchange of configuration information between educators working in different knowledge domains and institutions could help minimise the cost of this per OER. Use of the Pronunciation Lexicon Specification (Bagshaw, Burnett, Carter, & Scahill, 2007) could contribute towards this. The Pronunciation Lexicon Specification is an interoperable specification of pronunciation information for speech synthesizers and automatic speech recognition engines. The language is intended to be easy to use by developers while supporting the accurate specification of pronunciation information for international use.

- **IPR issues**
  Carey reports that alternative format production for library resources has never risen above 4% of standard-text publishing "this partly results from production methods but also from defensive copyright in which the rights of authors outweigh consumer access rights." (Carey, 2007). The Creative Commons licence adopted by OpenLearn explicitly permits remix [18], however other esources may have less clear licences.

5 Discussion and conclusions

In section 4 we put forward a number of factors we believe could affect the quality of learners' experiences with automatically generated open audio remixes. It is clear that further work is needed to establish the relative importance of these factors, and to discover other factors we may have missed. The Open University is planning to run a number of relevant studies during 2008 and 2009 as part of its Digital Audio Project (Doran, 2006). These studies are intended to investigate usability and accessibility issues concerned with digital audio educational resources in general and DAISY resources in particular. The studies are planned to include

- Lab based studies of a variety of learners (print-disabled and non-print-disabled) interacting with a variety of audio based educational resources, using a variety of players. These studies will be expected to yield quantitative and qualitative results.

- Longitudinal studies of a few learners using audio based educational resources as part of their daily routines. These will be expected to yield qualitative results, and will use techniques such as diary studies.

With respect to our subsidiary research questions (i.e. how can (semi-)automatic generation of audio remixes contribute to the sustainability of freely available OER? How can (semi-)automatic generation of audio remixes contribute to the provision of OER to consumers at no cost?) we identified a number of factors that will affect the impact of generation of audio remixes on the sustainability and cost of audio based OER. It is clear that further work is needed to establish what the key issues in the production process are. However, an investigation into the cost of configuring and programming a production process such as the pipeline should be carried out alongside research into the factors affecting the quality of learners' experiences.

Acknowledgements: For help, guidance and support - Robin Stenham, Doug Blane, Jonathan Fine, Peter Cox, Mary Taylor, Anne Jelfs, Chetz Colwell (OU), Karen Brasher (home), Linus Ericson, Martin Blomberg, Markus Gylling (Daisy pipeline developers) and others on the DAISY tech list. The initial stages of this work were supported by The William and Flora Hewlett Foundation as part of the OpenLearn Initiative.

6 References

American National Standards Institute. (2002). *ANSI/NISO Z39.86-2002 Specifications for the Digital Talking Book*. Retrieved 5/12/2007, from http://www.niso.org/standards/resources/Z39-86-2002.html

American National Standards Institute. (2005). *ANSI/NISO Z39.86 - 2005 Specifications for the Digital Talking Book*. Retrieved 31/5/2007, from http://www.niso.org/standards/index.html#Z39.86

Atkins, D. E., Brown, J. S., & Hammond, A. L. (2007). *A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities*. Retrieved 22/9/2008, 2008, from http://www.oerderves.org/wp-content/uploads/2007/03/a-review-of-the-open-educational-resources-oer-movement_final.pdf

Barrass, S. (2003). Sonification Design Patterns, *ICAD Proceedings*. Boston, USA: Internation Conference on Auditory Display.

Bellegarda, J. R. (2005). Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. *Pronunciation Modeling and Lexicon Adaptation, 46*(2), 140-152.

Brown, L. M., Brewster, S. A., Ramloll, S. A., Burton, R., & Riedel, B.-J. (2003). *Design guidelines for audio presentation of graphs and tables*. Paper presented at the 9th International Conference on Auditory Display (ICAD), Boston, Massachusetts.

Carey, K. (2007). The opportunities and challenges of the digital age: A blind user's perspective. *Library Trends, 55*(4), 767-784.

Cebeci, Z., & Tekdal, M. (2006). Using Podcasts as Audio Learning Objects, *Interdisciplinary Journal of E-Learning and Learning Objects* (Vol. 2, pp. 47-57).

DAISY consortium. (2007). *DAISY: Daisy Pipeline*. Retrieved 03/12/2007, 2007, from http://www.daisy.org/projects/pipeline/

DAISY Consortium. (2008). *DAISY: Requirements Gathering*. Retrieved 8/2/2008, from http://www.daisy.org/z3986/requirements/

Damper, R. I., Marchand, Y., Adamson, M. J., & Gustafson, K. (1999). Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. *Computer Speech and Language, 13*(2), 155-176.

Doran, C. (2006). Digital Audio (DA) Project Overview: The Open University.

Downes, S. (2007). Models for Sustainable Open Educational Resources, *Interdisciplinary Journal of Knowledge and Learning Objects* (Vol. 3).

Evans, C. (2008). The effectiveness of m-learning in the form of podcast revision lectures in higher education. *Computers & Education, 50*(2), 491-498.

Flowers, J. H. (2005). THIRTEEN YEARS OF REFLECTION ON AUDITORY GRAPHING: PROMISES, PITFALLS, AND POTENTIAL NEW DIRECTIONs, *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*. Limerick, Ireland.

Guillon, B., Monteiro, J.-L., Checoury, C., Archambault, D., & Burger, D. (2004). Towards an Integrated Publishing Chain for Accessible Multimodal Documents. In *Computers Helping People with Special Needs* (pp. 514-521). Berlin / Heidelberg: Springer

Hirst, T. (2007). *Feeding from open courseware: exploring the potential of open educational content delivery using RSS feeds.* Paper presented at the Proceedings of the OpenLearn2007 Conference, The Open University, Milton Keynes.

Laurillard, D. (2002). *Rethinking university teaching, a conversational framework for the effective use of learning technologies* (2nd ed.): RoutledgeFalmer.

Leith, L. (2006). *DAISY Intro Part 1: "Reading the DAISY Way"*. Retrieved 2/12/2007, 2007, from http://www.daisy.org/publications/docs/20070315155100/intro-article1.html

McAndrew, P., & Wilson, T. (2008). Pocketing the Difference: Joint Development of Open Educational Resources, *8th IEEE International Conference on Advanced Learning Technologies*. Santander, Cantabria, Spain.

Nes, M. E. S. (2007). *Appraising and Evaluating the Use of DAISY - For Print Disabled Students in Primary and Secondary Education.* UNIVERSITY OF OSLO, Oslo.

OpenLearn. (2008). *OU XML Schema improvements*. Retrieved 9/2/2008, from http://openlearn.open.ac.uk/mod/resource/view.php?id=277412

Phipps, L., Sutherland, A., & Seale, J. (2002). *Access All Areas: disability, technology and learning*. Retrieved 04/12/2007, 2007, from http://www.techdis.ac.uk/accessallareas/AAA.pdf

Reynolds, M. E., & Givens, J. (2001). Presentation rate in comprehension of natural and synthesized speech. *Perceptual and Motor Skills, 92*(3), 958-968.

Rowntree, D. (1994). *Teaching with audio in open and distance learning : an audio-print package for teachers and trainers*. London: Kogan Page.

Spiegel, M. F. (2002). Proper name pronunciations for speech technology applications. *Proceedings of the 2002 Ieee Workshop on Speech Synthesis*, 175-178.

Stephen Barrass and other contributors. (2003). *Sonification Design Patterns (wiki)*. Retrieved 9/2/2008, from http://c2.com/cgi-bin/wiki.pl?SonificationDesignPatterns

Taylor, M. (2004). *Guidelines for describing visual teaching material*. Retrieved 9/2/2007, from http://kn.open.ac.uk/public/workspace.cfm?workspacepageid=2709

Wittenbrink, H. (2005). *RSS and Atom*. Birmingham: Pakt Publishing Ltd.

7 Footnotes

[1] XHTML 1.0: http://www.w3.org/TR/xhtml1/

[2] The Open University Structured Authoring Generic Schema v0.6: http://labspace.open.ac.uk/file.php/1/common/OUGenericFull.xsd

[3] DAISY: About the DAISY Consortium http://www.daisy.org/about_us/index.shtml

[4] SMIL, Synchronized Multimedia Integration Language: http://www.w3.org/AudioVideo/

[5] MathML: http://www.w3.org/Math/

[6] AMIS: http://amis.sourceforge.net/

[7] Openlearn course units in DTB format: http://jime.open.ac.uk/2009/05/dtbs/

[8] Daisy Pipeline Transformer Description File (TDF) Grammar version 1.1: http://daisymfc.sourceforge.net/doc/developer/tdf-grammar-v1.1.html

[9] Daisy Pipeline Taskscript Grammar version 2.0: http://daisymfc.sourceforge.net/doc/developer/script-grammar-v2.0.html

[10] Pipeline Script: Narrator: http://daisymfc.sourceforge.net/doc/scripts/Narrator-DtbookToDaisy202.html

[11] The Narrator taskscript also creates a DTB conforming to an earlier version of the DTB standard; this can facilitate access to a book for users who do not have the latest reader software.

[12] Diversity and difference in communication: http://openlearn.open.ac.uk/course/view.php?id=1536&loginguest=true

[13] Dolphin EasyReader: http://www.yourdolphin.com/productdetail.asp?id=9

[14] PDF file available from page: http://openlearn.open.ac.uk/mod/resource/view.php?id=166541, by clicking link 'View document' i.e. http://openlearn.open.ac.uk/file.php/1536/K205_1%20Reader%20Chap%2012.pdf

[15] Fire Vox, a screen reader that is designed especially for Firefox: http://www.firevox.clcworld.net/

[16] MathPlayer, a MathML to speech player: http://www.dessci.com/en/products/mathplayer/

[17] RSS 2.0 Specification (RSS 2.0 at Harvard Law): http://cyber.law.harvard.edu/rss/rss.html , Atom Syndication Format 1.1 http://ietfreport.isoc.org/idref/draft-ietf-atompub-format/

[18] FAQ's - Intellectual Property http://www.open.ac.uk/openlearn/about-us/faq-ip.php

[19] A MIME type (or Multipurpose Internet Mail Extensions Media type) is a specification of a media type for use in Internet protocols such as e-mail and HTTP.

[20] XSL Transformations (XSLT) http://www.w3.org/TR/xslt