Chapter 3: Metadata

3.1 Introduction

3.1.1 Metadata is structured data that provides intelligence in support of more efficient operations on resources, such as preservation, reformatting, analysis, discovery and use. It operates at its best in a networked environment, but is still a necessity in any digital storage and preservation environment. Metadata instructs end-users (people and computerised programmes) about how the data are to be interpreted. Metadata is vital to the understanding, coherence and successful functioning of each and every encounter with the archived object at any point in its lifecycle and with any objects associated with or derived from it.

3.1.2 It will be helpful to think about metadata in functional terms as "schematized statements about resources: *schematized* because machine understandable, [as well as human readable]; *statements* because they involve a claim about the resource by a particular agent; *resource* because any identifiable object may have metadata associated with it" (Dempsey 2005). Such schematized (or encoded) statements (also referred to as metadata 'instances') may be very simple, a single Uniform Resource Identifier (URI), within a single pair of angle brackets < > as a container or wrapper and a namespace. Typically they may become highly evolved and modular, comprising many containers within containers, wrappers within wrappers, each drawing on a range of namespace schemas, and assembled at different stages of a workflow and over an extended period of time. It would be most unusual for one person to create in one session a definitive, complete metadata instance for any given digital object that stands for all time.

3.1.3 Regardless of how many versions of an audio file may be created over time, all significant properties of the file that has archival status must remain unchanged. This same principle applies to any metadata embedded in the object (see section 3.1.4 below). However, data about any object are changeable over time: new information is discovered, opinions and terminology change, contributors die and rights expire or are re-negotiated. It is therefore often advisable to keep audio files and all or some metadata files separate, establish appropriate links between them, and update the metadata as information and resources become available. Editing the metadata within a file is possible, though cumbersome, and does not scale up as an appropriate approach for larger collections. Consequently, the extent to which data is embedded in the files as well as in separate data management system will be determined by the size of the collection, the sophistication of the particular data management system, and the capabilities of the archive personnel.

3.1.4 Metadata may be integrated with the audio files and is in fact suggested as an acceptable solution for a small scale approach to digital storage systems (see section 7.4 Basic Metadata). The Broadcast Wave Format (BWF) standardized by the European Broadcasting Union (EBU), is an example of such audio metadata integration, which allows the storage of a limited number of descriptive data within the .wav file (see section 2.8 File Formats). One advantage of storing the metadata within the file is that it removes the risk of losing the link between metadata and the digital audio. The BWF format supports the acquisition of process metadata and many of the tools associated with that format can acquire the data and populate the appropriate part of the BEXT (broadcast extension) chunk. . The metadata might therefore include coding history, which is loosely defined in the BWF standard, and allows the documentation of the processes that lead to the creation of the digital audio object. This shares similarities with the event entity in PREMIS (see 3.5.2 , 3.7.3 and Fig.1). When digitizing from analogue sources the BEXT chunk can also be used to store qualitative information about the audio content. When creating a digital object from a digital source, such as DAT or CD, the BEXT chunk can be used to record errors that might have occurred in the encoding process.

A=<ANALOGUE> Information about the analogue sound signal path
A=<PCM> Information about the digital sound signal path
F=<48000, 441000, etc.> Sampling frequency [Hz]
W=<16, 18, 20, 22, 24, etc.> Word length [bits]
M=<mono, stereo, 2-channel> Mode
T=<free ASCII-text string> Text for comments
Coding History Field: BWF (http://www.ebu.ch/CMSimages/en/tec_text_r98-1999_tcm6-4709.pdf)

A=ANALOGUE, M=Stereo, T=Studer A820;SN1345;19.05;Reel;AMPEX 406
A=PCM, F=48000, W=24, M=Stereo, T=Apogee PSX-100;SN1516;RME DIGI96/8 Pro
A=PCM, F=48000, W=24, M=Stereo, T=WAV
A=PCM, F=48000, W=24, M=stereo, T=2006-02-20 File Parser brand name
A=PCM, F=48000, W=24, M=stereo, T=File Converter brand name 2006-02-20; 08:10:02

Fig. 1 National Library of Australia's interpretation of the coding history of an original reel converted to BWF using database and automated systems.

3.1.5 The Library of Congress has been working on formalising and expanding the various data chunks in the BWF file. *Embedded Metadata and Identifiers for Digital Audio Files and Objects: Recommendations for WAVE and BWF Files Today* is their latest draft available for comment at http://home.comcast.net/~cfle/AVdocs/Embed_Audio_081031.doc. AES X098C is another development in the documentation of process and digital provenance metadata.

3.1.6 There are however, many advantages to maintaining metadata and content separately, by employing, for instance a framework standard such as METS (Metadata Encoding and Transmission Standard see section 3.8 Structural Metadata — METS). Updating, maintaining and correcting metadata is much simpler in a separate metadata repository. Expanding the metadata fields so as to incorporate new requirements or information is only possible in an extensible, and separate, system, and creating a variety of new ways of sharing the information requires a separate repository to create metadata the can be used by those systems. For larger collections the burden of maintaining metadata only in the headers of the BWF file would be unsustainable. MPEG-7 requires that audio content and descriptive metadata are separated, though descriptions can be multiplexed with the content as alternating data segments.

3.1.7 It is of course possible to wrap a BWF file with a much more informed metadata, and providing the information kept in BWF is fixed and limited, this approach has the advantage of both approaches. Another example of integration is the tag metadata that needs to be present in access files so that a user may verify that the object downloaded or being streamed is the object that was sought and selected. ID3, the tag used in MP3 files to describe content information which is readily interpreted by most players, allows a minimum set of descriptive metadata. And METS itself has been investigated as a possible container for packaging metadata and content together, though the potential size of such documents suggests this may not be a viable option to pursue.

3.1.8 A general solution for separating the metadata from the contents (possibly with redundancy if the contents includes some metadata) is emerging from work being undertaken in several universities in liaison with major industrial suppliers such as SUN Microsystems, Hewlett-Packard and IBM. The concept is to always store the representation of one resource as two bundled files: one including the 'contents' and the other including the metadata associated to that content. The second file includes:

3.1.8.1 The list of identifiers according to all the involved rationales. It is in fact a series of "aliases" pertaining to the URN and the local representation of the resource (URL).

3.1.8.2 The technical metadata (bits per sample / sampling rate; accurate format definition; possibly the associated ontology).

3.1.8.3 The factual metadata (GPS coordinates / Universal time code / Serial number of the equipment / Operator / …).

3.1.8.4 The semantic metadata.

3.1.9 In summary, most systems will adopt a practical approach that allows metadata to be both embedded within files and maintained separately, establishing priorities (i.e. which is the primary source of information) and protocols (rules for maintaining the data) to ensure the integrity of the resource.

3.2 Production

3.2.1 The rest of this chapter assumes that in most cases the audio files and the metadata files will be created and managed separately. In which case, metadata production involves logistics — moving information, materials and services through a network cost-effectively. However, a small scale collection, or an archive in earlier stages of development, may find advantages in embedding metadata in BWF and selectively populating a subset of the information described below. If done carefully, and with due understanding of the standards and schemas discussed in this chapter, such an approach is sustainable and will be migrate-able to a fully implemented system as described below. Though a decision can be made by an archive to embed all or some metadata within the file headers, or to manage only some data separately, the information within this chapter will still inform this approach. (See also Chapter 7 Small Scale Approaches to Digital Storage Systems).

3.2.2 Until recently the producers of information about recordings either worked in a cataloguing team or in a technical team and their outputs seldom converged. Networked spaces blur historic demarcations. Needless to say, the embodiment of logistics in a successful workflow also requires the involvement of people who understand the workings and connectivity of networked spaces. Metadata production therefore involves close collaboration between audio technicians, Information Technology (IT) and subject specialists. It also requires attentive management working to a clearly stated strategy that can ensure workflows are sustainable and adaptable to the fast-evolving technologies and applications associated with metadata production.

3.2.3 Metadata is like interest — it accrues over time. If thorough, consistent metadata has been created, it is possible to predict this asset being used in an almost infinite number of new ways to meet the needs of many types of user, for multi-versioning, and for data mining. But the resources and intellectual and technical design issues involved in metadata development and management are not trivial. For example, some key issues that must be addressed by managers of any metadata system include:

3.2.3.1 Identifying which metadata schema or extension schemas should be applied in order to best meet the needs of the production teams, the repository itself and the users;

3.2.3.2 Deciding which aspects of metadata are essential for what they wish to achieve, and how granular they need each type of metadata to be. As metadata is produced for the long-term there will likely always be a trade-off between the costs of developing and managing metadata to meet current needs, and creating sufficient metadata that will serve future, perhaps unanticipated demands;

3.2.3.3 Ensuring that the metadata schemas being applied are the most current versions.

3.2.3.4 Interoperability is another factor; in the digital age, no archive is an island. In order to send content to another archive or agency successfully, there will need to be commonality of structure and syntax. This is the principle behind METS and BWF.

3.2.4 A measure of complexity is to be expected in a networked environment where responsibility for the successful management of data files is shared. Such complexity is only unmanageable, however, if we cling to old ways of working that evolved in the early days of computers in libraries and archives —