Unique identifiers for sound recordings

Introduction

The purpose of this document is to specify the file naming conventions for digital files held in the Sound Archive's collections, including in the Archival Sound Recordings Project.

For linkages to CADENSA and other electronic resources and including any future content management system that might be used in DOMS, the filename is the primary key to audio data and all associated files (e.g. metadata, edit lists, accompanying texts) therefore it is extremely important to the successful management of large digital collections that its structure is carefully defined. It is normal for computer system coding systems to have a fixed length coded identifier as the primary key. This offers the following advantages:

- Requires clearly defined rules for creating new UIDs
- Guarantees unambiguous recognition in the system (and for users who know the rules)
- Permits validation of the code or components of the code
- Supports wild card searching, sorting and reporting.
- Permits batch renaming if required for ingestion into different content management systems.

The structure of the UID (Unique Identifier) is derived from the solution that has already been employed by the Library in the *Collect Britain Project* and specified in the *Archival Sound Recordings* project, but the details are modified to allow variants of files and to improve readability.

The basis of the UID proposed here and used elsewhere (Collect Britain, ASR) is a 24-character alphanumeric file naming convention (excluding any .extension), designed to provide a unique filename for every audio file and any associated files of other types (images, texts). It has been designed to be able to accommodate existing, shelf-marking conventions employed within Library collections, and also to ensure the widest compatibility across different computer operating systems and storage media. It is also designed to be accommodated within the call number space allowed by the Unicorn software that is the platform for Library's Sound Archive on-line catalogue.

It conforms to ISO 9660 level 2 (i.e. a limit of 32 character length filenames) and uses the d-character set defined in ISO 646; i.e. the characters 'A'-'Z' (upper case), '0'-'9' (digits), '-' (hyphen), and '.' (full stop). It is further recommended that not more than one '.' be used. Lower case letters must not be used for the UID though it is accepted that in the case of a URI the domains (e.g. sounds.bl.uk) may be exclusively lower case.

Definitions and terminology

The Appendix 5 of the RFP introduces the specific meaning of six items *(represented here after in italics)*. Those definitions are copied below and new items have been introduced as to facilitate the reading of the specifications.

Carrier

The media a recording or file is stored on, for example a compact cassette, a coarse-groove 78 rpm disc or a minidisk.

Note: The carrier is limited by its ability to store files or folders. Each face of a microgroove is a carrier. The single layer of a DVD5 (4.7 GB) is one carrier. The two layers of a DVD9 (8.5 GB) is one carrier: at the point of view of the files and folders, the partition in layers is not visible)

Carrier-Stream

An assembled set of carriers. The set will be assembled from across all Content Packages. For example, course-groove and micro-groove records would make up one Carrier-Stream but be drawn from a variety of Content Packages.

Content Package

A thematic description of the various types of content to be included in this project. A 'Content package' is further split into 'works' and, for some of them, 'works' are further split into 'work-components'.

Note : In several places, the RFP expresses the fact that the 'content-packages' are excerpts from the 'British Library collections'.

Element

An element is each of the Data-Elements defined according to the "Dublin Core" methodology. In particular, each of the Data-Elements defined in the BLAP-S as use of the Qualified-Dublin-Core approach.

Index

The expression of a sequential link between items. For example, the two faces of a 33rpm microgroove (original carrier) could be indexed 01 and 02; the same index can be connected to each of the files representing a 'work-components' and those incremented to express the sequence of the 'work-components' to constitute the 'work'.

Original

Throughout this document, the word "original" as applied to sound recordings means the audio carrier provided by the Library to the Supplier, although strictly speaking it may not be "an original" recording, and the word "copy" will be used for the versions the Supplier will make.

Point

A 'location' in a stream: the location could be expressed in various ways (for example, time code, a tag or mark or chunk); metadata or synchronization could be connected to the point.

Resource Discovery

The data used to find resources from the catalogue or on the World Wide Web. This includes all the information required to retrieve accurate information about an object.

Segment

A part of a stream defined by two 'locations'; metadata or synchronization could be connected to the segment.

Shelf

One (or a set of) carrier physically bundled. The two faces (two carriers) of a microgroove constitute a 'shelf'.

Often, 'shelves' and 'carriers' are equivalent: a tape, for example.

Volume

A carrier holding specific content such as a set of files [and folders] in one [or various] formats, metadata, controls, … .

Volume-Stream

A chained set of 'volumes' such as all the 'volumes' holding all the representations coded in .mp3.

Work

A complete item such as the four movements, typically, of a string quartet or thematic breaks (or "chapters") within a lengthy oral history interview.

Wrapper

A technology for bundling metadata, files and folders pertaining to an item (logical and/or physical).

Work-Component

An elementary part of a 'work' such as each of the four movements, typically, of a string quartet or elementary thematic breaks within the chapters of a lengthy oral history interview.

UID structure

Each UID is made up of 8 separate 'fields' of information, each of which is represented by a fixed number of characters, as shown below.

e.g. the first ‘work-component’ of a ‘work’, derived (after digitization, restoration and encoding as a .WAV file) from an original typical coarse-groove disc in the BL collections (1CL0000237) will have the filename:

026A-1CL0000237XX-0001M0.WAV

This filename is constructed from the ten distinct sets of information as follows:

Resource ID	Document type	Separator	Root identifier	Separator	Part number	Start sequence	Document status	Version	Suffix
026	**A**	**-**	**1CL0000237XX**	**-**	**00**	**01**	**M**	**0**	**.WAV**
(resource_id)	(doc_type)	(r_separator)	(root_id)	(r_separator)	(stream_sq)	(compo_sq)	(doc_status)	(version)	(format)

Note: The names between brackets are possible refinements of the element “identifier” within the BLAP-S) [1]

Specifications for each information-set are given overleaf, and further examples of filenames are given at the end of this document.

[1] In the implementation at Memnon the names between brackets will be used to identify the elements constructing the names of the files. BL could also handle that as a refinement of the element “identifier” of the BLAP-S.

	Information set	Character type	No. of characters	Values	Explanation
1.	**Resource ID** (resource_id)	numeric	3	020 [Default, or unclassified curatorial area] 021 Oral history 022 Wildlife 023 Popular Music 024 Drama & Literature 025 World & Traditional Music 026 Classical Music 027 Soundscape 028 Broadcast 029 [undesignated]	refers to the originating resource, where 02 is the Library's Sound Archive and the last digit is one of the curatorial areas
2.	**Document type** (doc_type)	Alphabetic, or numeric upper case	1	A, I, T, V, R …. 0 to 9	a single letter code which identifies the type of document: A = Audio I = Photograph or other image T = Text V = video M = metadata E = edit list R = Relation [2] P = play-list X = combined documents (e.g. MPEG-4 with mix of images, video, metadata, audio, controls, etc) Duplicated **audio** files (several versions at different qualities or restoration techniques) will be coded by a digit (0 to 9). This is additional provision to (version)
3.	(r_separator)	HYPHEN	1		
4.	**Root identifier** (root_id)	Alphanumeric	12-	The "root_id" field is coded by a transposition of the value of the "shelfmark" (written and barcoded on the original product) according to the following rule. This field is always left justified, and padded with Xs up to a maximum of 12 characters and by replacing the non-alphanumeric characters by X	Derives from the shelf-mark for the particular item (or 'group' of items) in the collection. Other conventions are used for new recordings not accessioned or catalogued on CADENSA (see below). The separator separate the call number from the rest of the UID and improve readability Example: -1CL0000237XX-

[2] It is probable that some situation will occur where the BL will have to include Hyperlink files in the structure of the 'works' or of the 'originals'. If those hyperlinks are included in the METS file, there is no need of it but R is included as a provision.

				E.g.: 1CL0000237XX 1CDR0000237X C4052XXXXXXX C1078X7X4XXX is derived from C1078/7/4 For exceptions the transposition from the "shelfmark" to the "root_id" will be made by the BL.	
5.	(r_separator)	HYPHEN	1		
6.	**Part number** (stream_sq)	Numeric in some cases, Alphabetic in others	2	Always 2 characters	This field is intended to give the sequence of the volumes in their volume-streams or of the works in their work-streams having the same "root_id". Here are five representative cases : 1.- **The volume is not part of a stream**: the field is coded ZZ. 2.- **The Volume is part of a stream:** the field expresses the place of the volume in the sequence of volumes in the stream by the alphabetic coding: the first place is represented by AA, the second by AB, … up to YZ. 3.- **The work is not part of a stream:** the field is coded 00. 4.- **The work is part of a stream:** the field expresses the place of the work in the sequence of works in the volume-stream by the numeric coding: the first place is represented by 01, the second by 02, … up to 99. Always right justified & filled from the left with a zero. [3]

[3] *N.B. When the sequence of work-streams overlaps the sequence of a volume-stream, the "root_id" is derived from the volume carrying the start of the work. E.g. the string quartets 10, 12, 13 and 16 carried by three automatically-coupled LPs each having a specific "root_id" (1LP01138562X to 1LP0138564X) will be coded respectively*

- *1LP01138562X-01 (starts on disk 1, side 1, all bands)*
- *1LP01138563X-O1 (start on the disk 2, side 1, band 2)*
- *1LP01138564X-01 (starts on disk 3, side 2, all bands)*
- *1LP01138562X-02 (starts on disk 1, side 2, all bands).)*

					5.- **The special cases** are coded ZA to ZY (reserved for those inevitable, unforeseen cases) Comment: when there is a one-to-one mapping of the 'work' onto a 'volume', the files related to each original volume are marked with ZZ in the 'stream_sq' field and the files related to each work are marked with 00 in the 'stream_sq' field.
7.	**Start sequence** (compo_sq)	Alphabetic or numeric	2	Always 2 characters	This field is intended to give the sequence of the volumes-components or of the works-components having the same "root_id" and "stream_sq" Here are five representative cases : 1.- **The item is a 'volume'** (not a volume-component): the field is coded ZZ. 2.- **The Volume-Component is a part of a Volume.** This occurs when the digitization of a volume is segmented deliberately or by accident. Typical examples leading to this are: separation in bands or tracks in order to harmonize the audio level within the volume or other reasons concerning the restoration process; breaks in magnetic tape. The field expresses the place of the volume-component in the volumes by the alphabetic coding: the first place is represented by AA, the second by AB, … up to YZ . 3.- **The item is a 'Work'** (not a 'work-component'): the field is coded 00. 4.- **The Work-Component is part of a Work**. This occurs when the work is segmented. Typical examples are: movements of classical music compositions; the chapters in a reading of a book. The field expresses the place of the work-component in the work by the numeric coding: the first place is represented by 01, the

					second by 02, … up to 99. Always right justified & filled from the left with a zero. 5.- The **special cases** are coded ZA to ZY (reserve for unforeseen cases).
8.	**Document status** (doc_status)	alphabetic, upper case	1	M, P, A R,T,V	The status of the file. M = master (audio unaltered) [default]. P = playback (audio altered by noise reduction, filtering, editing etc). A = access (low resolution non-archival file prepared only for access, e.g. mp3 files for web). 'A' copies are created either from 'P' or 'M' copies. 'P' copies are created from 'M' copies. This character distinguishes files that are derivatives of each other and share the same sonic content and which would otherwise share the same name. It shows the relationship and archival role of the file. V = Validated used to indicate that a non-audio file is qualified for release R = auxiliary files used during the restoration process; those files are not planned to be deliverable T = temporary or trial files; those files are not planned to be deliverables. The Proxy .mp3 files are marked "T"
9.	**Version** (version)	Alphabetic or numeric	1	Any single digit or single character	To be used to distinguish the instances of files pertaining to the same object but issued for several reasons, like corrections of administrative or technical errors, improvement of the quality of the coding. Default is 0 for the M,P,A files (field 8) and is A for the R and T files (field 8).
10.	**Suffix** (format)	Alphanumeric, preceded by stop character. Letters in upper case	2,3 or 4 (excluding full stop)	Stop character, followed by any letters or numbers (usually 3 letters)	The file type, normally software-assigned when the file is created, e.g. .WAV, .TIFF, .TXT, .DOC, .WMA, .OGG, .RA, .MP3

Remark

The link between the disks, tapes and their sides and bands with the stream-sequence is not always obvious.

The **shelfmark** is connected to a set of disks or other physical products.
Most disks and tapes, by this definition, comprise more than one carrier, the two sides of an LP for example. Each side is a volume, part of a volume-stream. In the volume-stream, each part has an ordinal position reflected by the **stream-sequence** field.
The value of the “stream_sq” field of a specific volume is usually directly derived from the sequence of faces then the sequence of disks: Disk1 Side1 gives the “stream_sq” AA; Disk1 Side2 gives the “stream_sq” AB; Disk2 Side1 gives the “stream_sq” AC; Disk2 Side2 gives the “stream_sq” AD; and so on. However, some of the LP sets are organised to be listened in an automatic changers, for example for an opera. In that case, the “stream_sq” in the previous example becomes: Disk1 Side1 gives the “stream_sq” AA; Disk2 Side1 gives the “stream_sq” AB; Disk2 Side2 gives the “stream_sq” AC; Disk1 Side2 gives the “stream_sq” AD.

The **band sequence** (“band_sq”) within a side refines the expression of the sequence of the audio material.

Identifying objects

The identification of a **specific volume** in a set of volumes covered by the same shelfmark will be the concatenation of the “root_id” followed by the “separator” followed by the “stream_sq”.

For example: C 4 5 9 X Ø Ø 1 X Ø 1 X - A C

The identification of a **specific band** on a specific volume in a set of volumes covered by the same shelfmark will be the concatenation of the “root_id” followed by the “separator” followed by the “stream_sq” followed by the “comp_sq”.

For example: C 4 5 9 X Ø Ø 1 X Ø 1 X - A C A B

The identification of a **specific work** derived from one or several volumes (possibly within in a set of volumes covered by the same shelfmark) will be the concatenation of the “root_id” followed by the “separator” followed by the “stream_sq”.

For example: C 4 5 9 X Ø Ø 1 X Ø 1 X - 0 2

The identification of a **specific work-component** within a specific work will be the concatenation of the “root_id” followed by the “separator” followed by the “stream_sq” followed by the “comp_sq”.

For example: C 4 5 9 X Ø Ø 1 X Ø 1 X - 0 2 0 4

Identifying objects at Universal level (URI)

The identification of objects at world level with independence of the intended service will be expressed by the use of the concatenation of two fields.

The first field is the <domain-space>.

For audio related files and objects associated with the British Library the coding of the <domain-space> is as follows:

s o u n d s . b l . u k /

The coding of the second field depends on the type of resource:

For the ‘**works**’ and ‘**Volumes**’ the second field is expressed by the <root_id>, followed by the <r_separator>, followed by the <stream_sq>.

The URI of specific work becomes, in the preceding example:

sounds.bl.uk/C459X001X01X-02

For the ‘**works-components**’ the second field is expressed by the <root_id>, followed by the <r_separator>, followed by the <stream_sq>, followed by the <compo_sq>.

The URI of specific work becomes, in the preceding example:
sounds.bl.uk/C459X001X01X-0204

For the ‘**files**’, the second field is the file-name.

The URI of specific file of a work-component becomes, in the preceding example:
sounds.bl.uk/021A-C459X001X01X-0204A2.MP3

Examples of file names for the ASR project

In the ASR project, the Resource ID is allocated per Content Package :

- The Beethoven String Quartets are 026
- Wachsmann and Rycroft are both 025,
- The 400 Pop tracks are 023,
- The Visual Arts and Jazz in Britain are 021,
- African Writers will be 024,
- Soundscapes will be 027
- LBC Radio 028 (ditto).

Two examples analysed extensively will illustrate the coding in the case of the ASR project:
Note: To prevent confusion with letter 'O', all zeros are shown by the 'Ø' character

First example:

A specific cassette tape from the Visual Arts having the shelfmark C459/001/01 appears to be broken in the middle. They are five interviews on it.

resource ID	document type separator	Root Identifier	separator	Stream Sequence	Component Sequence	status	version	Format	comments
1 2 3 4 5 6 7 8 9	10	11 12 13 14 15 16 17	18	19	20	21	22 23 24		
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	A	Z	Z M Ø	. B W F	The intention was to clone the face 1 of the tape. The discovery that the tape is broken forces to prepare a fusion process.
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	A	A	A T Ø	. B W F	Same renamed
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	A	A	B T Ø	. B W F	Second part of the broken tape, face 1
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	A	Z	Z M 1	. B W F	The clone of the face 1 after fusion of the two previous files.
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	B	Z	Z M 1	. B W F	The clone of the face 2 after fusion of the two temporary files.
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	A	Z	Z T 1	. M P 3	The proxy of the face 1 of the tape.
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	A	B	Z	Z T 1	. M P 3	The proxy of the face 2 of the tape.
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	1	Ø	Ø R D	. W A V	The first interview under restoration 'work' (fourth restoration step)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	2	Ø	Ø R B	. W A V	The second interview under restoration as work' (second restoration step)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	1	Ø	Ø P Ø	. W A V	The first interview as a 'work' (Play-Back)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	1	Ø	Ø A Ø	. M P 3	The first interview as a 'work' (Access in MP3)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	1	Ø	Ø A Ø	. W M A	The first interview as a 'work' (Access in WMA)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	2	Ø	Ø P Ø	. W A V	The second interview as a 'work' (Play-Back)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	2	Ø	Ø A Ø	. M P 3	The second interview as a 'work' (Access in MP3)
Ø 2 1 A	-	C 4 5 9 X Ø Ø 1 X Ø 1 X	-	Ø	2	Ø	Ø A Ø	. W M A	The second interview as a 'work' (Access in WMA)

Second example:

A Box with the shelfmark 1LP0056669 containing three LP's includes four String Quartets of Beethoven. The Quartet n°10 occupies the disk 1 side 1 and continues on the disk 2 side 1, band 1.

resource ID	document type separator	Root Identifier	separator	Stream Sequence	Component Sequence	status	version	Format	comments
1 2 3	**4**	**5 6 7 8 9 10 11 12 13 14 15 16 17**	**18**	**19**	**20**	**21**	**22 23 24**		
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	A	Z	Z M Ø	. B W F	The clone of the face 1 of the first disk
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	B	Z	Z M Ø	. B W F	The clone of the face 2 of the first disk
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	C	Z	Z M Ø	. B W F	The clone of the face 1 of the second disk
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	D	Z	Z M 1	. B W F	The clone of the face 2 of the second disk
				…					
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	A	Z	Z T 1	. M P 3	The proxy of the face 1 of the first disk
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	B	Z	Z T 1	. M P 3	The proxy of the face 2 of the first disk
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	A	F	Z	Z R D	. W A V	The face 2 of the third disk under restoration (fourth restoration step)
Ø 2 6	E	- 1 L P Ø Ø 5 6 6 6 9 X X	-	Ø	1	Ø	Ø P Ø	. I P I	The Edit List of the 'work'
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	Ø	1	Ø	Ø P Ø	. W A V	The first quartet is a 'work', represented (as Play-Back copy, as one single file [with original version]). The quartet starts on the face 1 of the disk 1 (i.e. the first in the Volume 1 of the Volume-Stream)
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	Ø	4	Ø	Ø P 1	. W A V	The fourth quartet is a 'work', represented (as Play-Back copy, as one single file [with version 1])
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	Ø	4	Ø	2 P 3	. W A V	The second movement of the fourth quartet is a 'work-component', represented (as Play-Back copy, as one file [with version 3])
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	Ø	4	Ø	2 A 3	. M P 3	Same as .wav but access in MP3
Ø 2 6	A	- 1 L P Ø Ø 5 6 6 6 9 X X	-	Ø	4	Ø	2 A 3	. W M A	Same as .wav but access in WMA