Towards a Procedure for Quality Control on Large Collections of Digitized Audio Data: The Case of the "Fondazione Arena di Verona"

Federica Bressan[1] and Sergio Canazza[2]

[1] University of Verona, Department of Computer Science
Strada Le Grazie 15, 34134 Verona, Italy
federica.bressan_01@univr.it
[2] University of Padova, Department of Information Engineering
Via G. Gradenigo 6/b, 35131 Padova, Italy

Abstract. Audio recordings are an important documentary source for academic studies in many fields, from linguistics to anthropology. During the last decades, great efforts were made to develop guidelines and best practices for the preservation of audio documents, but not as much attention have been paid to quality controls on the re-mediation process and the output data.

This article presents the experience of the research project REVIVAL, aimed at the preservation of the audio documents stored in the archive of the Fondazione Arena di Verona, Italy, where a quality protocol was defined, and original software tools for automation of control were developed.

Keywords: Preservation, methodology, automatization, quality control.

1 Introduction

The importance of audio recordings (speech and music) as documentary sources for disciplines such as linguistics, musicology, ethnomusicology, anthropology and the like is fully recognized today. Accordingly, much efforts have been spent on the preservation of audio documents over the past decades (see [13] and [12] for an overview), and a variety of methodologies and best practices is currently made available by the international community (see [10,9,14,1]). However, the awareness that audio/visual carriers have an alarmingly short life expectancy compared to other pieces of cultural material, which can be measured in decades or years, caused a "rush" to digitization and an overall underestimation of the importance of quality controls during the process of *re-mediation* (the process of transferring the acoustic information from a medium onto another medium). This approach may have dramatic consequences on the authority of a document as a source for academic studies, besides invalidating the re-mediation process that needs to be repeated, which is not always possible as explained in Section 2.

This article presents the experience of the research project REVIVAL (REstoration of the VIcentini archive in Verona and its accessibility as an Audio e-Library), aimed at the preservation of the audio documents stored in the archive

C. Grana and R. Cucchiara (Eds.): MM4CH 2011, CCIS 247, pp. 103–113, 2012.

of the Fondazione Arena di Verona, Italy, with a special attention to the development of protocols and tools for quality control during the re-mediation process of audio documents.

Section 2 introduces the problem of unreliable data as output of unsafe preservation programmes. Section 3 presents the REVIVAL project, where algorithms for quality control and software tools, respectively presented in Sections 4 and 5, were developed.

2 The Threat of Unreliable Data

Poor methodologies for preservation may nullify the audio document as a source for scholarly studies or, worse, lay a doubt over analysis and theories that were based on those documents. In this sense, reliable repositories of documents should be a major concern for academics and specialized communities. They should be aware of the risks, and demand that preservation programmes are planned in order "to save history, not rewrite it" [5].

The process of re-mediation is fatally error prone (at technical, planning and operative level) and it often tends to indulge to the present aesthetical taste. These are factors that clearly motivate the definition of a strict protocol for the re-mediation of audio documents.

Secondly, preservation is not limited to "safeguarding the world's documentary heritage, [but it is aims at] democratizing access to it, and raising awareness of its significance and of the need to preserve it" [4]. In this sense, the creation of "preservation copies" (or "archive copies"), as defined in Subsection 4.1, is not enough: the data needs to be (unrestrictedly) accessed, meaning that tools for retrieval are needed, from low-level (locating and associating data) to high-level (Content Based Retrieval (CBR)).

Performing controls on digital data is not straightforward, especially for large collections. Producing invalid data during the re-mediation process means that the process needs to be repeated, which is not only time consuming, but may not be possible if the original carrier got physically damaged during playback, and it cannot be played again. The signal extraction from the original carrier is one of the most, if not the most, delicate step of the process, but all steps must be carried out with equal attention. Each step is closely related to the others, and early mistakes propagate in the workflow with ambiguous effects.

The preparation of the carrier (physical restoration) and playback allow for errors that damage the carrier directly, but the definition of the format for playback (e.g., speed, equalization curve and noise reduction system for analog recordings) is crucial, as wrong calibrations during the extraction of the signal invalidate the entire procedure.

What has been underestimated during the last few decades is the complexity of the audio re-mediation process, which does not coincide with simple A/D transfer, as is unfortunately often thought. In other words, in fact there are many different things can go wrong during the re-mediation process, and each step should always be performed in optimal conditions.

An additional reason for wanting quality data is that algorithms and tools for automatic classification and tagging, analysis and processing, generally perform better with quality data sets. Subsequently, it is convenient to control the process of producing data in order to feed these tools, which are being developed very fast, with good data sets.

3 The REVIVAL Project

REVIVAL (REstoration of the VIcentini archive in Verona and its accessibility as an Audio e-Library) is a national joint project between the Fondazione Arena di Verona and the Department of Computer Science of the University of Verona, with the scientific support of Eye-Tech[1]. It started in January 2009 and by the end of the second year, in December 2010, the partners agreed to extend it throughout 2011, basing on the quality of the objectives achieved that opened the way for further work.

Its main objective is the development of a HW/SW platform with the purpose of preserving and restoring and audio documents stored in the archive of the Fondazione Arena. The estimated value of the archive is 2,300,000 Euros. It comprises tens of thousands of audio documents stored on different carriers (from wax cylinders to digital carriers), hundreds of pieces of equipment for playback and recording (from wire to magnetic tape recorders and phonographs) and bibliographic publications (including monographs and all issues of more than sixty music journals from the 1940's to 1999). Along with a history of the recording techniques, the archive traces the evolution of a composite genre such as opera, with one of the largest collections of live and studio recordings in Italy. The most precious section of the archive is represented by the live recordings of the operas staged every year during the summer season at the Arena. The first opera festival was organized back in 1913 by the tenor Giovanni Zenatello and the theatre impresario Ottone Rovato, to celebrate the centenary of the birth of Giuseppe Verdi. Since 1936 the festival was organized by the "Ente Lirico Arena di Verona" (autonomous organization for lyrical productions), until the Ente Lirico was transformed into a private law foundation in 1998, the Fondazione Arena di Verona. The oldest recording available of the opera festival dates back to July 30th, 1968, and it is a performance of "Trovatore" by Giuseppe Verdi, featuring Leyla Gencer as Leonora and Carlo Bergonzi as Manrico. Figure 1 shows some of the open-reel tapes stored by the archive: all of them are unique copies. The archive is constantly growing with the new recordings of the current seasons, stored on HDD devices.

The first task of REVIVAL consisted in the development of an operational protocol aimed at the preservation of the audio documents stored by the archive [2]. The main international guidelines were considered ([14,1]) and trade-offs were made to meet the characteristics of the Arena archive, in terms of number and type of documents, genre of the recordings, objectives of the digitization. The documents for the re-mediation were selected according to the following criteria,

[1] http://www.eye-tech.it/

(a) Audio documents

(b) The laboratory

Fig. 1. Left (a): some of the open-reel tapes stored by the archive of the Fondazione Arena di Verona. All of them are unique copies. Right (b): a view of the laboratory that was set up withing the archive. Several tape recorders (in particular a Studer A-812 in the bottom right corner) and the A/D-D/A device can be seen. Between the loudspeakers for monitoring, on a lower shelf, the incubator for the thermal treatment of the tapes.

inspired by the (conflicting) classifications provided by the International Federation of Library Associations and Institutions (IFLA) [10] and the International Association of Sound and Audiovisual Archives (IASA) [9]:

1. original carriers in immediate risk / on endangered media;
2. documents in regular demand;
3. documents which depend on obsolete or commercially unsupported systems.

Figure 1 shows a view of the laboratory, which is equipped to fully support the preservation process, from the restoration of the physical carriers (e.g., thermal treatment in incubator) to the A/D-D/D transfer, from metadata extraction to multimedia processing. With the experience matured on the number of documents digitized, some experiments were planned 1) to have a better understanding on how magnetic tapes respond to thermal treatment (commonly used to compensate for the sticky shed syndrome [7]), in collaboration with the Department of Mathematical, Physical and Natural Science of the University of Verona and the Department of Chemical and Process Engineering of the University of Padova, and consequently plan better treatments; and 2) to find out if the audio signal obtained from a tape played in the intended direction (sound forward) equals the one obtained with the same tape running the opposite direction (sound backwards). The rationale for this experiment is that extracting the signal in one playback session for tapes showing two independent mono tracks using a stereo head (then splitting the stereo signal and reversing one channel by means of a wave editor) would save half of the time employed to read the tape (which may last up to 4 hours) and would prevent fragile carriers from being played back twice.

4 The Re-mediation Process

Two types of preservation can be distinguished for audio documents: passive, meant to defend the carrier from external agents without altering its structure, and active, which involves information transfer onto new media. A protocol for the re-mediation of audio documents consists in the formalization of the steps for active preservation.

The purpose of this protocol is to ensure that the loss of information during the re-mediation process is minimized. A number of different tasks are necessary to ensure that *all* information is read/generated, interpreted, represented and described adequately. The required input is: 1) the original document; 2) knowledge about the recording technique; 3) knowledge about the history of the recording. The output is the preservation copy, as defined in Subsection 4.1.

The process is structured in three steps (before playback, playback and after playback), each of which is articulated in procedures and sub-procedures. The output of each procedure and sub-procedure is either data, a report or a different state of the system.

The first level is general and applies to all types of carriers. The second must be occasionally adapted to the type of carrier that is being treated, and the third is completely carrier type dependent. Figure 2 shows the general scheme of the re-mediation process.

The re-mediation process requires a combination of conceptual, technical and also manual skills that result in a complex professional profile. Besides, during the process things can go wrong in very many ways. Therefore, each procedure was divided into simple tasks, described by a separate workflow, and each block is extensively commented. Exceptions are managed, and the precision of the

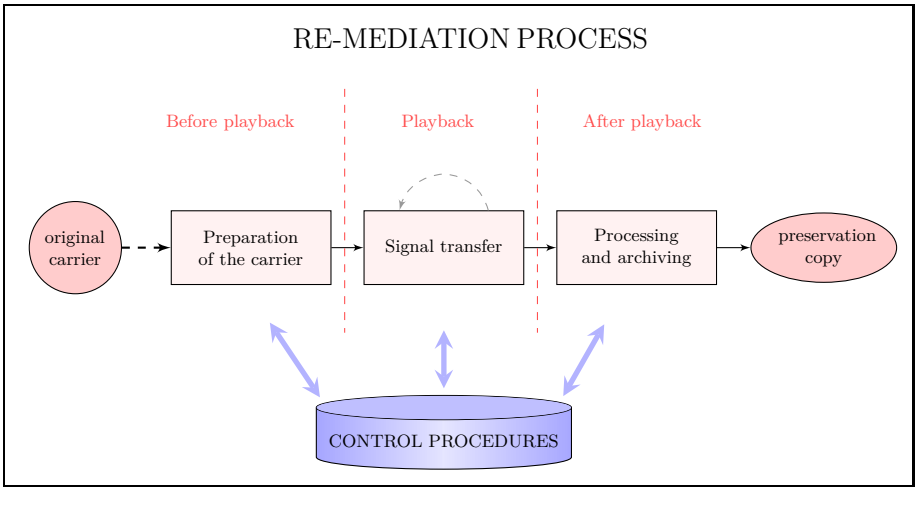

Fig. 2. General scheme of the re-mediation process, from the original carrier to the preservation copy

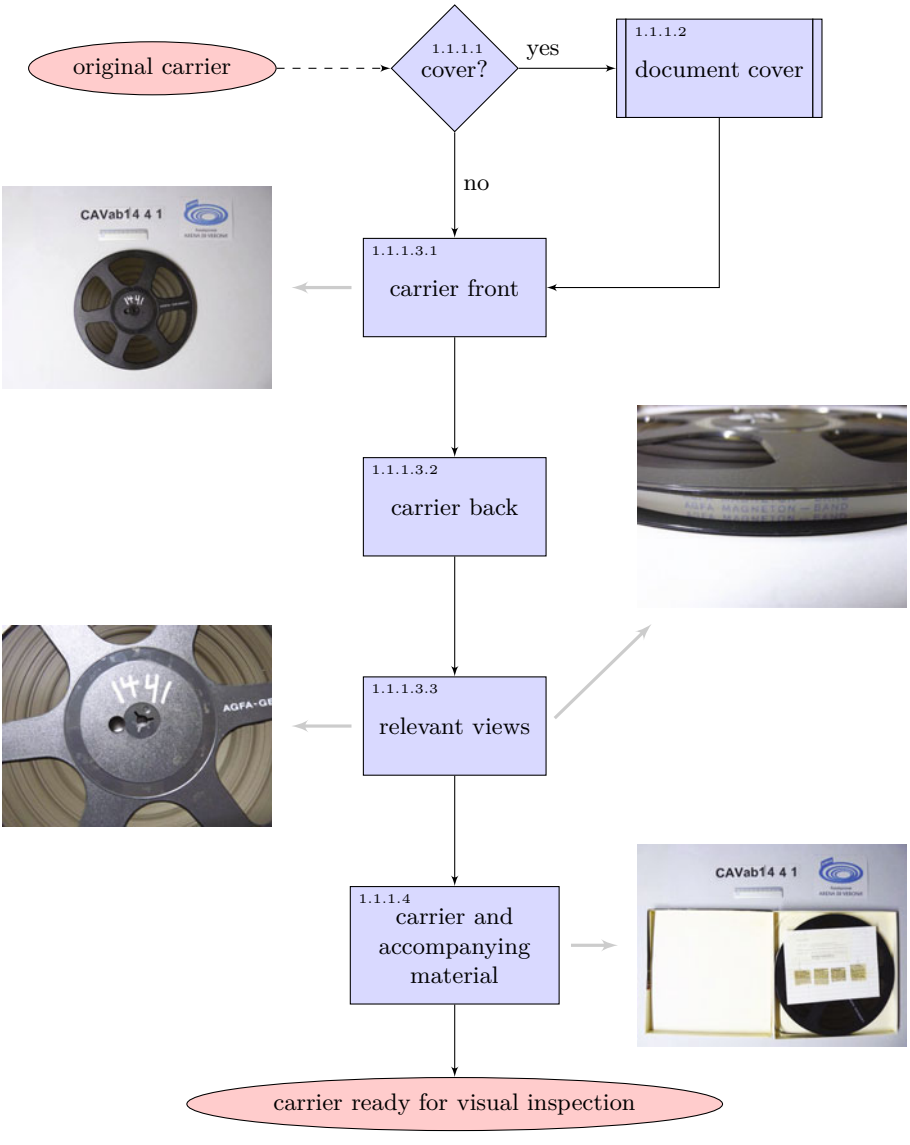

Fig. 3. Example of flowchart, describing procedure 1.1.1 (Preparation of the carrier → Physical documentation → Pictures of the document) mentioned in Section 3. Blocks marked with double lines (such as 1.1.1.2) are sub-functions with a separate description. Each block is extensively commented and exceptions are managed.

descriptions is as rich as possible in order to reduce the indecision that comes from the large variety of carriers and the numberless combinations of symptoms presented by the carriers.

Figure 3 provides an example of a very simple workflow: it represents the step 1.1.1 of the process (Preparation of the carrier → Physical documentation → Pictures of the document). In the notation adopted by the project reports, blocks marked with double lines are sub-functions described separately.

The structure of the workflow in Figure 3 is straightforward, but the example is representative because the aim of the document is to provide precise descriptions of each task. To achieve this goal, visual material and notes are associated to the blocks, with several references to separate sections where more material is presented and commented. This workflow describes the physical documentation of the carrier, and it comes with a section where guidelines for visual documentation of cultural heritage are presented [6], along with suggestions for the setup of a photographic workspace and a number of warning and tips that make a difference in the quality of the output data.

If all of the workflows reach the end successfully, the re-mediation process is complete. The expected output is a preservation copy of the original document, of which a general description is provided in the next Subsection.

4.1 Preservation Copy

The preservation copy (or Archive copy) is defined in [8] as the "artifact designated to be stored and maintained as the preservation master. . . . Such a designation means that the item is used only under exceptional circumstances." In concrete terms, the preservation copy is a data set that groups all the information carried by the original document. This is not limited to the audio signal,

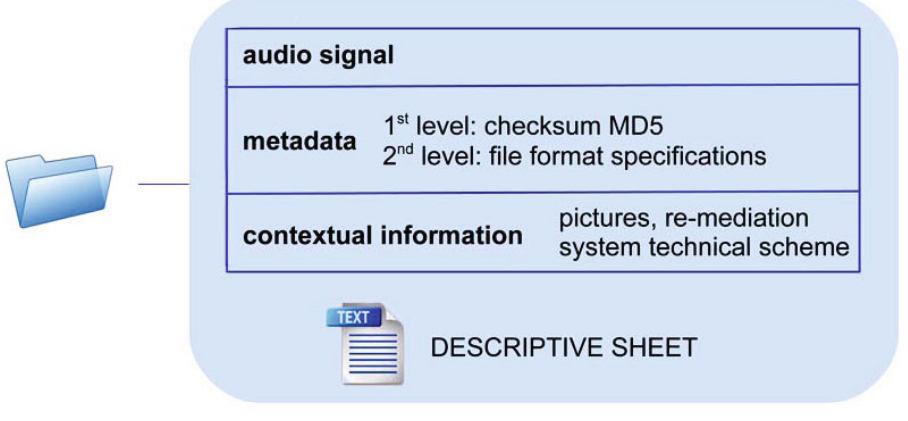

Fig. 4. Logical representation of the elements contained in a preservation copy

but it comprises metadata and contextual information[2], that is a complete documentation of the physical carrier and the accompanying material, plus a set of metadata that are are produced during the re-mediation process. Figure 4 shows the logical representation of a preservation copy. For further descriptions and background, see [3].

5 Automation

A quality re-mediation of audio documents requires specialized infrastructures, multi-disciplinary trained personnel for preservation and long-term maintenance, resulting in a system of resources that not all archives in the real world are able or willing to afford. This entails that many institutions may start digitization campaigns with technological makeshifts and inadequate methodologies, resulting in invalid documentary corpora. To avoid this risk, it is important to make cultural actors aware of the consequences of an uncontrolled use of technology, and thus to encourage them to maximize the quality of their preservation programme, according to the international standards and best practices.

Where applicable, automation is often a good solution to the problem of reducing costs. Moreover, automation brings several benefits that go beyond simple task delegation. It allows to minimize mistakes, perform (cross-)controls, and it allows people to concentrate on higher level tasks with better attention.

In this sense, automation can be of great help in preservation programmes both at local and national scale. However, it is to be noted that there are some crucial steps in the re-mediation process that are likely to remain refractory to full automation, such as the analysis and handling of the original carrier before playback. Nevertheless, some tools supporting semi-automation will be mentioned later in this paragraph.

In this context, automation mainly applies to:

1. procedures and tasks of the re-mediation process;
2. large sets of data consisting in audio and metadata (i.e., output of remediation process).

During the REVIVAL project, some utilities have been developed to perform controls over the entire collection of preservation copies and to automatize time-consuming and low-level tasks that are intrinsic in any archival routine.

The utilities were developed in Java [3], because i) the workstations of the archive mount different OS and Java allows cross-platform compatibility as well as high-level abstraction from the physical machine, and ii) fast code development and a large availability of libraries are necessary to design, implement and test the tools during the short life-cycles of a project.

[2] In this context, the term "metadata" refers to the content-dependent information that can be automatically extracted from the audio signal, and "contextual information" to the additional content-independent information.

[3] Java Version 1.6.0_24.

Although some of the tasks implemented by the utilities are not audio specific, meaning that some of them could be performed with a generic piece of free software, they got integrated in a tool especially designed for audio digital archives with the consequence that the level of automation got increased. At the same time, path variables and serialized objects have been kept as general as possible, making it easy for other archives to benefit from these tools in preservation programmes based on a similar approaches.

The personnel of the Fondazione Arena was asked for feedback during the development of each utility, defining the tasks that needed to be automated or controlled out of the real work in the laboratory. Another concern was that the utilities would be easy to use for people with little or no computer skills, which is often the case of archive personnel. Each piece of software was provided a GUI (an example is shown in Figure 5) and it can be launched by clicking on an icon as most desktop applications.

1. Utility to perform controls on the entire collection of preservation copies, searching for: empty directories, missing directories, anomalous directories, mismatch in file names and file formats, missing checksums.
2. Utility to rename audio and visual data pertaining to a preservation copy (based on a drag-and-drop interface).
3. Utility to perform a control on the checksums of the entire collection of preservation copies and calculate the missing ones (grouped in a single file

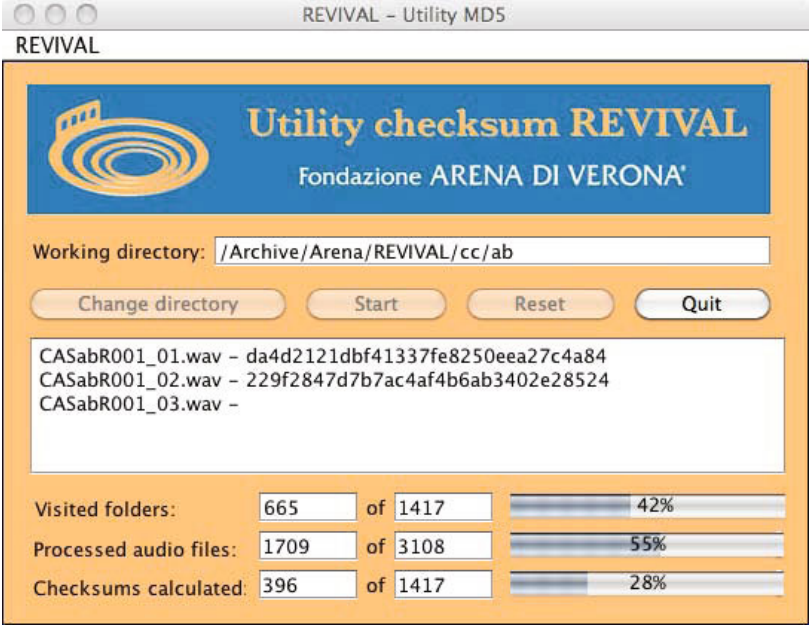

Fig. 5. One of the software tools developed for the Fondazione Arena di Verona

according to the structure of the preservation copy). Figure 5 shows this utility at work.
4. Utility for the long-term maintenance of the archive: it re-calculates the checksum of the audio files in each preservation copy and confronts it with the existing one.

Future work includes the integration of the functionalities described above into an independent work panel, in order to assist the personnel all along the re-mediation process. Besides the benefits of automation mentioned at the beginning of this Section, the work panel would ensure that procedures are carried out according to the protocol discussed in Section 4, and it would help dealing with problems and exceptions. The work panel would be able to extract and insert the data directly from a database, allowing i) additional cross-controls over the audio archive and the database; and ii) decreasing the possibility of introducing mistakes when a new record is created. A prototype implementing JHOVE[4] for metadata extraction is currently under test.

6 Conclusions

This article described the process of re-mediation for audio documents and pointed out the reasons why it is necessary to define procedures and perform controls on the process of A/D-D/D transfer and the subsequent management of digitized/digital data. In particular, the experience of the REVIVAL project was presented, with a description of the adopted methodology and the software tools that were developed to perform controls on the collection of preservation copies and to automatize time-consuming and low-level tasks have been described. The personnel of the archive of the Fondazione Arena di Verona got involved in each step of the process, and the software tools were progressively integrated in the archival routine. These tools will still be used after the end of the project, thus ensuring that the scientific protocol is maintained in the future.

References

1. Boston, G.: Safeguarding the Documentary Heritage. A guide to Standards, Recommended Practices and Reference Literature Related to the Preservation of Documents of all kinds. UNESCO (1988)
2. Bressan, F., Canazza, S., Salvati, D.: The vicentini sound archive of the arena di verona foundation: A preservation and restoration project. In: Workshop on Exploring Musical Information Spaces (WEMIS) in conjunction with ECDL 2009, Corfu, Greece, pp. 1–6 (October 2009)
3. Canazza, S., Orcalli, A.: Preserving musical cultural heritage at mirage. Journal of New Music Research 30(4), 365–374 (2001)
4. Edmonson, R.: Memory of the World: General Guidelines to Safeguard Documentary Heritage. UNESCO (February 2002)

[4] JHOVE is a Java based application developed by JSTOR and the Harvard University Library for the identification, validation and characterization of digital objects [11].

5. Edmonson, R.: Audiovisual Archiving: Philosophy and Principles. UNESCO, Paris, France (April 2004)
6. Galasso, R., Giffi, E.: La documentazione fotografica delle schede di catalogo - metodologie e tecniche di ripresa. Tech. rep., Istituto Centrale per il Catalogo e la Documentazione (ICCD) - Ministero per i Beni e le Attività Culturali (1998)
7. Hess, R.: Tape degradation factors and challenges in predicting tape life. ARSC Journal 39(2), 240–274 (2008)
8. IASA: The IASA Cataloguing Rules. IASA Editorial Group (1999)
9. IASA-TC 03: The safeguarding of the audio heritage: Ethics, principles and preservation strategy. Tech. rep. (2005)
10. IFLA - Audiovisual and Multimedia Section: Guidelines for digitization projects: for collections and holdings in the public domain, particularly those held by libraries and archives. Tech. rep., International Federation of Library Associations and Institutions (IFLA) (March 2002)
11. JSTOR, the Harvard University Library: Jhove – jstor/harvard object validation environment, `http://hul.harvard.edu/jhove/`
12. Orcalli, A.: On the methodologies of audio restoration. Journal of New Music Research 30(4), 307–322 (2001)
13. Orio, N., Snidaro, L., Canazza, S., Foresti, G.L.: Methodologies and tools for audio digital archives. International Journal on Digital Libraries 10, 201–220 (2009)
14. Storm, W.D.: The establishment of international re-recording standards. Phonographic Bulletin 27, 5–12 (1980)