THE SOUND OF AN ARTICLE: PRESERVING AUDIO AND VIDEO SUPPLEMENTAL MATERIALS FROM PRINT JOURNALS

Jonathan Ponder, ITHAKA-JSTOR, Content Management, USA

1. Introduction

This paper addresses issues around the conversion and digital preservation of audio and video content issued as supplemental material to print journal issues, and specifically audio and video originally issued on physical media such as CDs, as opposed to born-digital media files. I am not going to address the processes and technical details of actually doing this conversion and digital preservation work, mostly because, at JSTOR,[1] we have not actually done much work with this material yet. Instead, I'm going to deal with 1) a precise description of the particular subset of content that I am concerned with, 2) the story of it at JSTOR and how it has come to my attention in my work, 3) a little bit about the larger context of journal supplementary material and digital preservation, and 4) some of the challenges and questions this audio-visual content on physical media raises. I am thinking about this presentation and paper as a way to simply alert the reader to this content, discuss some of the issues surrounding it, and check in to see if this sounds like an area of concern to others like it does to me. Along the way, I will share what we have done (but again, mostly not done) at JSTOR to convert, preserve, and provide access to this kind of supplementary journal content, and I will hopefully also communicate the value, benefit, and interest of this kind of material and why we should care about preserving it in the context of its related journal articles.

2. What is JSTOR?

JSTOR, ITHAKA S+R, Portico, and Artstor are all products and services of ITHAKA.[2] ITHAKA is a non-profit organization whose stated mission is to "work[] with the global higher education community to advance and preserve knowledge and to improve teaching and learning through the use of digital technologies." JSTOR helps fulfill part of this overall mission. JSTOR started in 1995 with scanning and providing electronic access to the full back runs of academic journals, as well as committing to the long term preservation of the journals that it signs agreements with. JSTOR currently provides access to thousands of journals, books, and primary sources from over 1,000 different publishers, and it serves 9,000 universities, schools, and institutions and 6 million unique monthly users. JSTOR has been an important player from the beginning in the movement toward digitization and digital preservation of academic journals.[3]

3. Journal audio-visual media collection

After I had been at JSTOR for a little while as a Metadata Librarian, I became aware of "the box of random stuff". This box was filled with a various assortment of material that had come to JSTOR over the years along with print journals (either with donation issues or subscription issues). The large majority of the items were CDs, DVDs, and CD-ROMs, but there was also microfiche, cassette tapes, obsolete computer diskettes, a few vinyl records, and a pair of 3-D glasses, among other things.

1 JSTOR is pronounced JAY-stor. A shortened form of "journal storage", "JSTOR" was a nickname used during its initial conception and founding and then became the actual name. More about what JSTOR is in a moment.

2 The name ITHAKA looks like an acronym but is not; it is the name of the organization. ITHAKA S+R is ITHAKA Surveys and Research, a service that "provides intelligence about emerging technologies to guide the transformation of higher education." Portico is "a community-supported digital archive [that] preserves over 350,000 e-books and e-journals for future scholars." Artstor derived the spelling of its name from JSTOR and is a digital repository for art that "provides 2+ million high-quality images and digital asset management software to enhance scholarship and teaching." Details and quoted descriptions in this note come from "ITHAKA: Our Work," ***ITHAKA.org***, accessed December 8, 2016, http://www.ithaka.org/our_work.

3 Details and quoted descriptions of ITHAKA and its services in this paragraph come from the ITHAKA website: "ITHAKA: Our Mission," ***ITHAKA.org***, accessed December 8, 2016, http://www.ithaka.org/content/our-mission; "ITHAKA: Our Work," ***ITHAKA.org***, accessed December 8, 2016, http://www.ithaka.org/our_work.

The inventory and acquisitions staff were responsible for this material, but there was no written inventory or control over what all was actually there. I was only aware of the "box of stuff" when, in my work, I would sometimes encounter a piece of media and ask inventory what to do with it. They would point me to the "box of stuff," the contents of which we all hoped would someday be able to be made accessible on the JSTOR website. Before I started working at JSTOR, there was a time when the production staff inserted explanatory pages into articles with associated media, and even scans of the media in certain cases, to alert users to its existence, even though we did not have the actual content accessible online. This was not happening any more by the time I started at JSTOR. We were just continuing to collect the material with the hope of using it at some point. I became interested in this material and volunteered to take responsibility for it. I went through and made a spreadsheet inventory of all of it, connecting each piece of media to the articles or issues that they are associated with.

4. Overview of content in collection

So what did I find in this random collection of "stuff?"

- **Sound recordings.** There are sound recordings that contain music for the most part but also poetry and other types of sound recordings, such as field recordings and experimental recordings. One example is a collection of poets reading their work that was compiled and released by *The Hudson Review* with their 55th Anniversary Issue.[4] Another example is *Computer Music Journal*,[5] which puts out a CD or DVD with every volume that is a curated and theme-based compilation of recordings around some topic related to the volume and its articles. Some *Computer Music Journal* anthologies contain both sound and video where both the audio and video are the focus. There is, for example, a video for Vol. 30 that shows videos of a robot flutist and videos of other types of robots playing music. Videos such as these demonstrate the workings of these machines that are created by researchers as well as the music that the machines are able to create.
- **Videos.** Videos in the collection include stand-alone works of film, video illustrations of work discussed in articles (like the robot flutist), samples of videos that are raw data from research studies, or videos of papers/presentations from a conference. In a few cases there are videos called "Video Papers" where the video constitutes the substance of a journal article and there is not a textual counterpart. We have several examples of this in JSTOR, for example, one where the online table of contents for the issue lists the video papers but there is not actually any content because the video is not yet accessible on the website and there is no text version of the "paper".[6]
- **Multimedia Presentations.** Some examples of multimedia presentations in our collection include:
  - Videos of conference presentations that also include the text of the paper and/or presentation slides, all embedded on a DVD or other software program.
  - A CD-ROM that supplemented an issue of *Memoir (Society of Vertebrate Paleontology)*[7] and contains an interactive interface that navigates a 3-D digital atlas of an alligator skull ("Alligator: Digital Atlas of the Skull").
- **Records and cassettes.** These again contain sound recordings but are on media that present an additional layer to digital conversion and preservation work.
- **Computer discs.** There are a few different formats of old computer discs where I have not been able to find out what is on them.

4 *The Hudson Review* 56, no. 1 (Spring 2003), accessed December 8, 2016, http://hudsonreview.com/issue/spring-2003/.
5 *Computer Music Journal*, accessed December 8, 2016, http://www.mitpressjournals.org/cmj.
6 An example: Ferdinando Arzarello and Ornella Robutti, "Approaching Functions Through Motions Experiments," *Educational Studies in Mathematics* 57, no. 3 (2004), accessed December 8, 2016, http://www.jstor.org/stable/4150294
7 Rowe, et al., 1999

- Other collection details
  - Appendix-like data that supports research in articles, either in digital form or on microfiche. When this kind of data is in some kind of reproducible page-like form, we have added it as page images to articles, but when it is data either embedded in database software or in some kind of program, we have not done anything with it.
  - Image files corresponding to articles that were either not included in the print version of the article or were present in black and white while often the digital file is in color and/or in high resolution.
  - Digital surrogates of print articles or issues in PDF form on a CD. These are not of much interest unless they contain color or higher resolution images than the original print version.
  - There is a total of 236 separate items to date, but we continue to get more.
  - There are lots of different file types, both standard and proprietary.
  - Most of the time, material relates to specific articles, but it may relate to the issue as a whole or to the theme of a theme issue.
  - These materials may or may not be referenced in some way in the print version of articles or issues.

5. The larger context

In discussing this topic, I want to make a clear differentiation between *born-digital* audio-visual content in the networked environment and audio-visual content on *physical media*, whether digital or analog. Supplementary journal content issued on physical media is a particular subset of content that is related to but not exactly the same as supplementary content in the born-digital networked environment. It is a clear difference that people understand, but I have also found that, as I bring up this topic with different people, they sometimes do not understand the distinction clearly enough. Someone may think they know what I'm talking about and may say, "Oh yes, this has been addressed or is being addressed by this or that project." The issue of supplementary journal article content has been addressed and worked on to a certain extent in the born-digital environment with various projects and initiatives:

- NISO Recommended Practices for Online Supplemental Journal Article Materials[8]
- The RMap Project, which is an initiative by the Data Conservancy, Portico, and IEEE "to make it possible to preserve the many-to-many complex relationships among scholarly publications and their underlying data…."[9]
- A LOCKSS paper in *Information Standards Quarterly* titled "Archiving Supplemental Materials"[10]

Projects and reports like these have dealt with the digital preservation of supplemental content, but these projects assume electronic formats that are in the networked environment. I am concerned with the transition from the print and physical media environment to the digital networked environment: audio, video, and multimedia content that was issued originally on physical media but now needs to be *converted* and *preserved* in the networked environment. This particular subset is not considered in other projects or initiatives as far as I can determine.

Supplementary journal content on physical media is not a new problem; it has been one for a long time. In the world of print periodicals in libraries, extra pieces of media that came along with issues of print journals did not fit the library model of shelving and binding journals. Libraries often have not known what to do with this material and have found places apart from the journal issues and articles to hold the media, or they have left the media stuffed into the journals on the shelves where their existence is largely unknown and uncontrolled.

8 "NISO/NFAIS Supplemental Journal Article Materials Project," ***NISO.org***, accessed December 8, 2016, http://www.niso.org/workrooms/supplemental

9 "RMap Project," ***rmap-project.info***, accessed December 8, 2016, http://rmap-project.info/rmap/

10 Rosenthal and Reich, 2010

There is now an opportunity in the digital world to bring articles and related audio-visual content together in a way that wasn't possible before, and I would say that these materials are valuable, beneficial, and worth being preserved for what they can bring to research and learning. But preserving them and bringing them together with their articles is not automatic. As academic journals move increasingly to born-digital publication and as back issues of print journals are increasingly converted to digital versions and libraries de-accession print copies, there is a risk of these supplemental materials on physical media being lost and separated for good from their journal issues and articles. There are a number of challenges that I have faced in trying to work with this kind of material, and I will discuss these in the next part of this paper.

6. Challenges and issues

6.1 Organizational priority and resource issues

The allocation of organizational priorities and resources is probably the key challenge that relates to all of the others because, if JSTOR were to make this content a priority and to decide to put resources toward it, then we would be better able to deal with the other challenges, such as digitization, more effectively. These challenges would still raise questions and issues to work through, but they would be less daunting because there would be people and time to work on them. There is really a small amount of this audio-visual content in comparison to the rest of JSTOR's content and other work. Everyone agrees that it is interesting material and that it would be ideal to have it converted, preserved, and accessible to users, but on the other hand, users are not asking for it or complaining that it is missing. Probably almost nobody—including publishers, users, and most JSTOR staff—is aware that it exists and is missing from the publicly available content, so getting the time and resources to do something with it might be difficult.

I decided to start working on a small scale to try to push this work forward. I talked to our Content Development Unit about the possibility of beginning to approach publishers and explore rights issues for just one or two journals at a time and, in this way, take on a manageable amount of work. In this way, I might also begin to learn what kind of questions will come up and how much work will be required. I sent my colleague on JSTOR's Content Development team a list of some of the journals that have sound recordings for her to choose a few publishers who may be responsive and easy to work with on this. The one response so far ended up being very positive and it looks like this content for this publisher will be easy to deal with, at least as far as rights issues. The publisher for two journals that have yearly collections of sound recordings and videos informed us that they have the rights to permit the work on those discs to be included in any database in which the journals are held.

6.2 Unclear rights issues

Now, more about rights issues. This is often the first question that comes up when we start talking about working on this content. Do we have the rights to convert, preserve, and make it accessible? Does the publisher whom we deal with own the rights? This media content often has stated copyright holders that are different than the journal publisher or current journal rights holder that JSTOR has negotiated with, therefore it is not clear and cannot be assumed that the journal that issued the content has rights to it beyond its original issue. In compilations, there are usually multiple individual rights holders for the different works.

- Examples:
  - Annual recording compilations put out by *Computer Music Journal* and *Leonardo Music Journal*,[11] for example, have a copyright statement for the compilations as a whole but also a different copyright statement for each of the individual creators/contributors of each track.
  - It is a similar situation with the recordings of poets reading their work for the *Hudson Review* that was mentioned earlier. The compilation is copyrighted by the journal, but the individual authors still hold the copyright to their poems.

This looked like a real problem to me: the possibility of having to seek out numerous individual rights holders for every individual work on each different disc and the amount of time, work, and resources that would require. I was happily surprised that our first attempt looks to be simple and straightforward. Hopefully this will be the case for the others, but there may be more complicated cases to address.

In JSTOR's case, for a long time, the rights for these supplemental materials were not discussed or even considered in negotiations for digitizing the print issues of journals, simply because nobody thought about them or was aware that they would be published with the print journals. As I understand it, JSTOR's Content Development team now addresses this topic in negotiations, but for the first 15 years or so of JSTOR it was not addressed. So for this earlier content, JSTOR will need to go to the publisher or rights holder and ask specifically about the rights for the audio-visual supplemental content.

These rights issues have, unfortunately, caused a dynamic between different departments in a way that relates to organizational priorities and resources. Production and development don't want to put work into converting the material and developing systems and platform code to handle it until we know that the rights issues are cleared. On the other hand, Content Development, who deals with the publishers, is hesitant to approach publishers and create expectations about this material on the public site until they know it's actually going to happen according to some kind of timeline. This dilemma is one reason that I started trying the approach of just beginning to explore a little bit at a time.

6.3 Assessing the nature and importance of the material

Another question that needs to be addressed in this undertaking is assessing the nature and importance of this material, not the intrinsic value and importance of the content itself but rather of this particular copy that we have in this context. Is its issue with the journal that is in JSTOR unique? Is the journal the only place where this content was ever issued, and maybe the only place where it exists, so that it needs to be preserved with the journal, or is it also available and archived elsewhere? Examples of these contrasting situations might include, for example, recordings issued with *The Quarterly Journal of the Library of Congress*,[12] which are likely preserved by the Library of Congress vs. the recordings gathered and issued by the *Hudson Review* of poets reading their work, which probably only exists in this particular CD issue. The *Hudson Review* recording liner notes describe how they requested these readings and recordings from the poets and that they were recorded in different ways in many locations. Then the recordings were sent to the journal especially to be placed on this CD.

A further question is, if a particular work *is* archived elsewhere, is it necessary to preserve it in the context of the journal? Is it integral to understanding the article, or is it truly supplemental to an article or issue? This is a distinction that is addressed by the NISO Recommended Practices for Online Supplemental Materials, cited earlier. This document makes a distinction between content that is supplemental to an article but not needed to understand the article vs. content that is truly integral to understanding an article and

11 *Leonardo Music Journal*, accessed December 8, 2016, http://www.leonardo.info/lmj/about.html.
12 Shirley, 1981

therefore must be present with the textual article. Of course, my bias is that, regardless of whether it is supplemental or integral or whether it is preserved elsewhere or not, if it was issued with the journal, having it preserved and accessible in that context will enrich the experience or the arguments of the article and is worthwhile to preserve within that context. But in the context of organizational priorities and resources, I may need to address some of these questions and be able to defend the value of using time and resources to preserve and provide access to this material. Do we *need* to preserve it in JSTOR? If we don't, will it still be preserved and accessible somewhere else?

6.4 Assessing how much we don't have

Another challenge or question that arises is knowing how much is out in the world that we don't have or don't know about, and given that, what should our collection policy be for this content? Our collection of this material has always been and still remains passive. The original model for JSTOR didn't expect this kind of content. We just started getting it sometimes, either with donated, loaned, or subscription issues. Especially for donated or loaned back issues, there is probably no way to know how much of this type of content might have been issued originally but didn't make it to us for one reason or another when we acquired and scanned the journals.

As an example, when I looked further into the Vol. 30 DVD for *Computer Music Journal* (mentioned earlier), I discovered that this particular DVD disc includes material relating to Vol. 30, but it is also a reissue of all of the recordings issued from Vol. 4 (1980) to Vol. 14 (1990). These earlier recordings were originally issued on thin vinyl "soundsheets" bound as inserts into the pages of the journal, and we do not have these vinyl soundsheets for this journal. We apparently did have at least some of them because one of them was scanned and placed as a page within the relevant article, but we no longer had them by the time I started working with the collection. I also discovered looking into the Vol. 30 disc that there were CDs issued from Vol. 19-Vol. 21 that we do not have. Based on this example from just one journal, it is clear that there might be much out there that we don't have and don't know about.

So then the question comes up, what is our policy going to be? And what is our responsibility? If we begin to preserve *some* of this content for *some* journals, are we then going to begin to search for it? The answer to that from those making decisions about budgets and resources is likely, "No, probably not." But it does raise the question and concern that, if we preserve this content for some journals, will there be an expectation, and maybe some responsibility, for us to preserve and provide access whenever it exists with JSTOR journals and to consider it missing journal content if we don't have it? We do know that some libraries make the decision to de-accession their print copies of a journal based on JSTOR acquiring that journal. When they do that, are they aware of this audio-visual content and are they de-accessioning that too? If they are aware of it, are they assuming that JSTOR has it as well as the articles? Regardless of what libraries may or may not be aware of, when de-accessioning happens there will be fewer and fewer copies of the audio and video content saved and accessible for a given journal. Given this, maybe JSTOR should assume some responsibility to make sure that this kind of content is preserved for the journals it commits to preserve.

6.5 Conversion and obsolescence issues

One last issue is that of conversion and obsolescence of the media containing this content. There are costs involved in converting this material, and the processes get more complicated and the costs increase as the media or file types get older and become obsolete. The physical media such as records, cassette tapes, and old computer disks require an extra layer of work to be converted to digital files. Multimedia presentations are another difficult type; they are usually on some kind of proprietary software that in some cases works and opens up easily but in other cases does not and will require further work and exploration to determine what is on the media. Once the actual content is determined, there is the question and difficulty of converting, preserving, and maintaining these kinds of complicated digital files in a sustainable way. With costs and more complicated processes, we face the problem of being able to allocate

resources, but then, at the same time, if we wait too long, we may not always be able to recover some content from physical media or file types if they degrade or get so old that there is not hardware or software upon which to run them. I am well aware that this tension between the costs of converting and the risk of not converting needs to be managed and dealt with.

7. Conclusion

I will conclude with one final thought about organizational priorities and resources and about the importance (or not) of preserving this content with its associated journals. I mentioned before that nobody is asking for this content or complaining that it is missing. If it is, in fact, important for JSTOR to preserve this material, it would be good for us to hear from you or from other users of JSTOR. I would be interested to hear from anyone about this, and it would also be good for JSTOR to hear directly from users who are looking for content like this that is missing. There are arguments to be made about preserving it because it's the "right" thing to do, however, it would be more effective for JSTOR to hear directly from the users that they are missing this material and that they want it to be accessible. JSTOR could then set priorities and aim resources toward the problem. Please let me and let JSTOR know what you think about this. I believe that these materials are interesting and important, and that they can enrich learning and research when present. My plan is to keep pushing forward and hopefully to be able to navigate the challenges and get this audio-visual content preserved for and accessible to students and researchers.

8. Postscript

I received feedback from a few people who attended my presentation at the 2016 IASA Conference. In general, there seemed to be agreement both that this is a topic of interest to the IASA community and also one that is difficult to address. Gene DeAnna, Head of the Recorded Sound Section of the Library of Congress, introduced the papers at my session, and he commented to me at the time and in a subsequent email exchange that this kind of material does present challenges at the Library of Congress. Currently, it is separated from the print content that it originally accompanied, and although its existence is noted in cataloging records and the media is being conserved in good storage for the Library, the separation into entirely different buildings and locations creates an impediment to access. He concurred that some of the biggest challenges are 1) finding the resources to digitize and preserve this content and 2) assessing which of the content is unique and more critical to preserve and which is preserved elsewhere and so not as critical to preserve. I have subsequently learned in additional email exchanges with another librarian at the Library of Congress that there is a digital preservation program for this kind of physical or tangible media and its content at the Library, but I have not been able to determine yet how far back this preservation plan goes and whether it covers journal supplementary material going back many years or only to when the preservation plan started and going forward (DeAnna and Leigh, 2016).

A different attendee from the Library of Congress expressed to me his opinion that this kind of audio-visual supplementary content is indeed crucial and that it should definitely be part of the mission of an organization such as JSTOR to preserve it along with the journals it commits to preserve. Another attendee suggested to me that I look at the work that has been done with scientific datasets for ideas and help on how to deal with the content I am concerned with. Acknowledging the distinction between born-digital supplementary content and supplementary content on physical media, he said that there still is likely some good information in what others have started to do with supplementary content to guide me in my work on this collection.

References

Arzarello, Ferdinando, Ornella Robutti. (2004). "Approaching Functions Through Motions Experiments." *Educational Studies in Mathematics* 57, no. 3. Accessed December 8, 2016, http://www.jstor.org/stable/4150294.

Computer Music Journal. (2016). Accessed December 8, 2016. http://www.mitpressjournals.org/cmj.

DeAnna, Gene and Andrea Leigh. (2016). 'Follow up to IASA presentation on audio/video journal content.' email thread with author, 2016.

The Hudson Review 56, no. 1 (Spring 2003). Accessed December 8, 2016. http://hudsonreview.com/issue/spring-2003/.

"ITHAKA: Our Mission." (2016). *ITHAKA.org*. Accessed December 8, 2016. http://www.ithaka.org/content/our-mission.

"ITHAKA: Our Work." (2016). *ITHAKA.org*. Accessed December 8, 2016. http://www.ithaka.org/our_work.

Leonardo Music Journal.(2016). Accessed December 8, 2016. http://www.leonardo.info/lmj/about.html.

"NISO/NFAIS Supplemental Journal Article Materials Project." (2016). *NISO.org*. Accessed December 8, 2016. http://www.niso.org/workrooms/supplemental.

"RMap Project." (2016). *rmap-project.info*. Accessed December 8, 2016. http://rmap-project.info/rmap/.

Rosenthal, David S.H., Victoria Reich. (2010). "Archiving Supplemental Materials." *Information Standards Quarterly* 22, no. 3 (Summer 2010). doi:10.3789/isqv22n3.2010.04.

Rowe, Timothy, Christopher A. Brochu, Matthew Colbert, John W. Merck, Jr., Kyoko Kishi, Emin Saglamer and Sheila Warren. (1999). "Introduction to 'Alligator: Digital Atlas of the Skull.'" *Journal of Vertebrate Paleontology* Vol. 6 (June 14, 1999): 1-8. Stable URL: http://www.jstor.org/stable/3889339.

Shirley, Wayne D. (1981). "Reconciliation on Catfish Row: Bess, Serena, and the Short Score of Porgy and Bess." *The Quarterly Journal of the Library of Congress* 38, no. 3 (Summer 1981): 144-165. Stable URL: http://www.jstor.org/stable/29781900