

International Association of Sound
and Audiovisual Archives



Internationale Vereinigung der
Schall- und audiovisuellen Archive



Association Internationale d'Archives
Sonores et Audiovisuelles



Asociación Internacional de Archivos
Sonoros y Audiovisuales



iasa

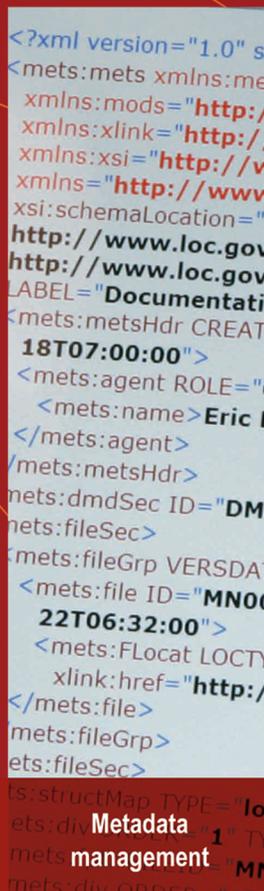
iasa journal

no 38, January 2012

www.iasa-web.org

YOUR MEDIA ARCHIVING PARTNER

SERVICES AND SOLUTIONS



Large scale audio & video digitization
Pre-cataloguing and indexation automation tools
Consulting for audiovisual archiving project

Memnon Archiving Services SA
27 rue du Belvédère - 1050 Brussels - Belgium- info@memnon.eu
www.memnon.eu

2011 Conference Notice	2
Editorial	3
President's Letter	5
Articles	
Digital sense and nonsense: digital decision making in sound and audiovisual collections <i>Ute Schwens, Deutsche National Bibliothek (German National Library)</i>	8
"And the survey says...OK, but the data says!" <i>Cassandra Gallegos, George Blood Audio and Video</i>	20
Simplifying semantics: Linked Data rules OK! <i>Fran Alexander, BBC Information and Archives</i>	27
Map Your Voice: a user-generated archive of English accents <i>Richard Ranft, Head of Sound & Vision, The British Library</i>	35
Tape dehydration as part of the "Journale" Project: on dealing with sticky-shed syndrome <i>Stefan Kaltseis and Anton Hubauer, Österreichische Mediathek</i>	40
Storage Strategy Tools <i>Matthew Addis, Mariusz Jacyno, Martin Hall-May, IT Innovation Centre, University of Southampton, UK, and Richard Wright, Archive Research, BBC Research and Development, UK</i>	46
Fonoteca Nacional of México launches the Platform for Sound and Audiovisual Heritage Preservation of Iberoamérica <i>Álvaro Hegewisch, Fonoteca Nacional (IASA membership vice-president) and Xilonen Luna, Comisión Nacional para el Desarrollo de los Pueblos Indígenas</i>	54

**43RD ANNUAL CONFERENCE OF THE INTERNATIONAL
ASSOCIATION OF SOUND AND AUDIOVISUAL ARCHIVES (IASA)**

IN TRANSITION: ACCESS FOR ALL

**New Delhi, India
6 – 11 October 2012**

On behalf of the **AIIS Archives and Research Centre for Ethnomusicology**, we invite you to New Delhi for the 43rd IASA Annual Conference.

It is a great honour for us and our hosting partners to have you in India for IASA 2012. We are certain that this will be a very stimulating and interesting conference, with the opportunity of meeting old friends as we do at all IASA conferences, but in a location that is quite different, getting to know India through new friends that we hope you will make, and interacting with new challenges.

This will be a very important event for those of us who work in the field of audio visual archiving in India, and we will do our best to make this not only professionally exciting but to have a great time!

We believe that the theme of our 2012 Conference is relevant to all of us, the wider IASA membership and our situation in India. We are at a crossroads (and have been for a while!). It is a period of transition with which we in India as in many other countries are struggling with.

New technologies, changing standards, challenges of digitization, finding financial support for preservation etc. The aim is of course 'Access for all' and that is the biggest challenge.

As we work towards the conference in the weeks and months ahead we will be finalizing the themes for the conference.

Please see the Call for Papers at:
<http://2012.iasa-web.org/call-presentations>

Please find all conference information on the conference website at:
<http://www.2012.iasa-web.org/>

For any further information or questions please contact the Organising Committee and the conference administrator through enquiries@iasa-conference.com

**International Association of Sound
and Audiovisual Archives**



**Internationale Vereinigung der
Schall- und audiovisuellen Archive**



**Association Internationale d'Archives
Sonores et Audiovisuelles**



**Asociación Internacional de Archivos
Sonoros y Audiovisuales**



Dear Readers,

Herewith another issue of the IASA Journal packed with knowledge and shared experience. Our last conference, held in Frankfurt in September 2011, carried the title “Digital sense and nonsense: digital decision making in sound and audiovisual collections”. Our keynote speaker, Ute Schwens (director of the German National Library, one of the hosts of the conference), gave a very clearly laid out justification for the need for a digitization strategy at the highest level. Her paper is included here in full.

Similarly, though in an entirely different context, we learn from Álvaro Hegewisch (Fonoteca Nacional de México and new IASA membership vice-president) and Xilonen Luna (Comisión Nacional para el Desarrollo de los Pueblos Indígenas, México) of attempts at the national level to introduce and disseminate standards and best practices for sound and audiovisual archiving across a country and a region, using the knowledge and expertise within IASA. Keeping at that high level, we get advice on decision-making on storage strategies from world leaders, Matthew Addis, Mariusz Jacyno, Martin Hall-May (IT Innovation Centre, University of Southampton) and Richard Wright (BBC Research and Development). And finally, we have Cassandra Gallegos’ paper highlighting the importance of not just collecting but really reading and *listening* to what data collected about our collections is telling us so that we can make better informed decisions.

At project level, Stefan Kaltseis and Anton Hubauer’s (Österreichische Mediathek) paper on tape dehydration, as opposed to baking, offers a practical and cheap solution to the, well, *sticky* question of sticky shed syndrome. When this paper was presented at the conference in Frankfurt, there was much discussion and debate. It struck me that this is an area we could usefully continue to debate and share ideas and experiences on. I hope this paper helps to kick that discussion off.

And once again we have a project presentation from Richard Ranft (British Library). While this presents another of the Library’s attempts at public engagement and crowdsourcing, it fits with our theme of making sense and nonsense by exploring the meaning of these attempts and how they assist us as archivists in making sense of our collections.

Finally, in IASA Journal 37, we included a cornerstone article that introduced the potential future of our Cataloguing Committee, now the Organising Knowledge Task Force. Guy Maréchal and Simon Rooks presented the theory and principles of the Semantic Web. In this issue Fran Alexander (Taxonomy Manager, BBC Information and Archives) furthers our understanding by simplifying, explaining terminologies and providing practical examples of engaging with those theories and principles. I, for one, have found this very helpful. I hope you will too.

2011 was an election year and you will see from our President’s letter in the following pages that there have been some changes to the Executive Board. Of relevance here is that we have a new editor. Cassandra Gallegos holds a Master of Science in Information Studies as well as a Certificate of Advanced Study in Preservation Administration. She has worked in the digitisation department at the Harry Ransom Humanities Research Center in Austin, Texas and at George Blood Audio and Video, where she has been concerned with digitisation standards and best practices. IASA will be in good hands with Cassandra as editor.

This means that I bow out with this issue, which has been treated as a hand-over issue. Cassandra will be all set to go it alone on issue 39 due out in June/July 2012. I’d ask you all to assist her by sending in your suggestions and contributions for the journal. I wish Cassandra all the very best with this challenging and rewarding position and I wish you all informative and happy reading into the future!

Yours truly,
Janet Topp Fargion
Editor IASA (Out-going)

Dear Readers,

I am both excited and honoured to be taking over as editor of the IASA journal. Over the past few months, I've observed the excellent work that Janet does in preparation for the journal and if I perform these duties half as well as she does then I am confident that my time as editor will be successful. I look forward to working with and learning from all the members of the IASA executive board as well from the IASA membership. I attended my first IASA conference this past September in Frankfurt. Never before have I met such a welcoming and intelligent group of people. Thank you all for making me feel as if I truly belong in the world of audio archiving.

Sincerely,
Cassandra Gallegos
Editor IASA (In-coming)

Time of transition and consolidation...

This is a year of change for our association — the proposed constitutional change, the theme of our next conference. But immediately, I am of course thinking of the change as a result of an election: an election year which brings a significant change of officers to the IASA Board. I am very happy to be part of this team that bears a healthy balance of fresh new members with new eyes and ideas with continuing members bringing into the transition experience and stability. Judging from the enthusiastic atmosphere of our first board meeting, this promises to be a great team.

The new IASA executive board includes the returning members Pio Pellizzari as Vice President and convenor of the training Committee, Richard Ranft continuing to shoulder our presence online as Web Editor, Kevin Bradley, returning in the role of Past President and myself as brand new President. New to the IASA executive are Cassandra Gallegos, from George Blood Audio and Video to shoulder the task of editor, Marit Hamre from the NRK to take on the role of Treasurer, Alvaro Hegewisch, from the Fonoteca Nacional in México as Vice President with responsibility for membership, Bruce Gordon from the Eda Kuhn Loeb Music Library at Harvard as Vice President for conferences, and Lynn Johnson from e.tv in South Africa as Secretary-General.

I would like to thank outgoing board members Dafydd Pritchard who administered the membership, Janet Topp Fargion who flagged the editorship and most of all Anke Leenings for her exemplary management of the association's accounts as one of the longest standing board member in the history of the association. The new board is also thankful for their continued support to the new members of the board during this transition.

“In Transition” is not only one of the themes for this coming board period, but also the theme for our next conference in the fantastic location of New Delhi, India, putting the focus on ‘access for all’. By the time you read this you will also have seen the Call for Papers, so let me invite you to send in a proposal for a presentation by the end of January 2012 and help make this conference the best ever!

In my inaugural President's speech I said that we live in a world of constant change, and these changes are happening at an ever increasing rate: national boundaries have changed, the political atmosphere here and there are as volatile as ever, the economic situation is ever so unstable with repercussions felt by each and every one of us, not to mention the ever more rapid development of the digital era, where devices now become obsolete almost before you get your hands on them. Change is especially hard to tackle for archivists like us who naturally prefer to preserve things as they are rather than having to deal with change.

In this unstable world, it is a source of reassurance to know that some things don't change and that we, through IASA and its reliable IASA-ness, have a stable core to hold on to and rely upon. A common denominator for IASA-ites seems to be a passion for our profession, an enthusiasm to hear of each other's experiences and a generosity in sharing our knowledge for the greater benefit of the archiving community.

To me that is synonymous with the brilliant theme of the Sydney Conference: “No Archive is an Island” — my institution is not alone with its challenges, and indeed I, in my new role as President, am definitely not alone, surrounded as I am by a wealth of competency, experience and enthusiasm for our cause.

We are not alone in dealing with the ever changing technical world.

One of the jewels of IASA, no doubt, is our Technical Committee, and over the past few years we've seen the publication of the TC-04 “Guidelines on the Production and Preservation of Digital Audio Objects” which has, as you know, become a work of reference in most sound

and audiovisual communities and is already in its second edition. The TC-05 on storage is well on its way, we've made great progress on the TC-06 on video and we have received ideas for further publications to consider.

Our TC consists of a prolific community of experts drawn, of course, mainly from among our members. But its phenomenal strength, I believe, lies in a quality typical of IASA-ness, namely its openness, looking beyond the collective knowledge of its members for the competency necessary to complete its work, welcoming views, sharing, debating and finding solutions — in brief — working in a way that is a great benefit to IASA and its standing in the AV-world as well as an example in how we indeed can work *together* for the greater benefit.

We are not alone as an association.

Over the last years, we've kept up and improved on our relations with other organizations, taking a lead through our very own Kurt Deggeller in the CCAA (Coordinating Council of Audiovisual Archives Associations). Keeping communication channels open with our sister organizations, we've been out there offering audiovisual sessions, workshops and tutorials at the ICA, at IFLA and in cooperation with other bodies. We are also monitoring several international projects by being there and waving the AV-flag.

We are not alone within our association.

We've improved on the setup of our committees and sections day during the conferences to have them perceived as an integral part of the conference program by encouraging relevant papers to be read at each meeting. This is where we meet others with the same dilemmas, or find a forum for understanding our special issues, etc. The IASA-ness is found here as well — for example by transforming the Cataloging Committee into a Knowledge Management Task Force to tackle the challenges of changed digital premises.

We are not alone, but part of an ever healthier membership. In 2005, at the beginning of my board tenure, I was handed a responsibility for the membership. You may remember me talking of gardening a membership, which was like pruning a rosebush by handing the scissors to the neighbor, then giving it personal and urgent attention for a stronger re-growth. We did cut about a third of the non-financial membership, but by gentle coaxing and being present at other conferences and workshops, the membership grew back to its original number well within that board period. Now this membership is a paying and more active membership — something that benefits IASA with a healthier bank account and fresh input from our newer members.

So, during the last two board terms in which I've served as Vice President, we've taken steps to adapt our organization to a changing world — many of these processes are still ongoing and we're still 'in transition' so to speak:

- We've cleaned up the membership, yes, but we now have the challenge of maintaining and expanding that membership. We're going to look at what it takes to keep members and attract new ones, how to keep our doors open to aspiring members and to show that membership is an investment that brings tangible fruits.
- We're in the process of assessing the needs of the previously named Cataloging Committee — the OK (Organizing Knowledge) task force is working with that.
- We're looking at implementing the running of conferences more within IASA ranks to make it less daunting for venues to take on a hosting function.
- I've already named all the publications that are under way — contributors to these need to have the room to work and the processes in place to materialize and become available.
- We are in the middle of a constitutional change that needs to be completed. If the membership agrees with the proposals, we are looking at incorporating the association, with the changes that this requires in the conducting of IASA-business.

- 
- We are now also looking at how to meet the challenges of board's work — both in work load and in financial resources to be able to carry out the work and meetings — so that it becomes both possible and attractive to serve as officers of the board.
 - We are now also in a much healthier financial state, and we will be looking at ways to finally address the wish list of projects.

Change will continue, premises will change, money will appear and disappear. Change is unavoidable and it easily brings us out of our comfort zone, but it is only through change that we can improve.

With all these ongoing changes and the adapting processes we've started on so far, I think it's important for this board period to facilitate the changes and prepare the fertile ground to allow the important processes already underway to come to fruition and stabilize within this board term. I would therefore call this board period a period of transition and consolidation, a time to bolt these changes into place and become part of the core of our association, part of our IASA-ness, so to speak. This does not mean, of course, that we can't at the same time, work to keep the ground fertile for further growth with our heads up and eyes open, ready to tackle the next challenges that are bound to come along.

So the red thread in this association of ours, of which I, and I am sure all of you are so proud, is the sense of community, an open and welcoming community with a stable core of IASA-ness, able to weather the storms of a changing world. With such a supportive team of colleagues and close collaborators, I am certainly looking forward to navigating these waters in the coming three years. Indeed thank you for your continuing support!

Jacqueline von Arb
Stavanger, Norway
November 2011

DIGITAL SENSE AND NONSENSE: DIGITAL DECISION MAKING IN SOUND AND AUDIOVISUAL COLLECTIONS

Ute Schwens, Deutsche National Bibliothek (German National Library)

Keynote speech at the 42nd IASA annual conference in Frankfurt, September 3rd to 8th, 2011

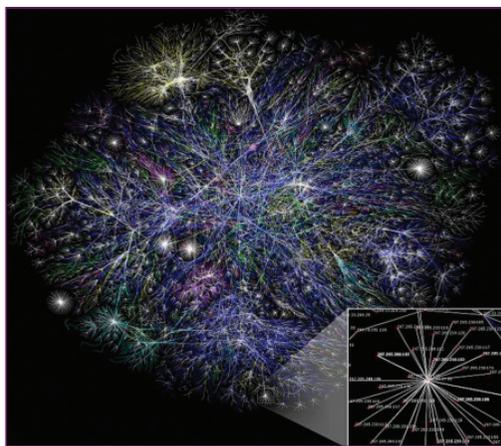


Figure 1: Visualization from the Opte Project of the various routes through a portion of the Internet

“The Internet is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide. It is a network of networks that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic, wireless and optical networking technologies. The Internet can also be defined as a worldwide interconnection of computers and computer networks that facilitate the sharing or exchange of information among users. The Internet carries a vast range of information resources and services, such as the inter-linked hypertext documents of the World Wide Web (www) and the infrastructure to support electronic mail.

Most traditional communications media including telephone, music, film, and television are reshaped or redefined by the Internet, giving birth to new services such as Voice over Internet Protocol (VoIP) and Internet Protocol Television (IPTV).

Newspaper, book and other print publishing are adapting to Web site technology, or are reshaped into blogging and web feeds. The Internet has enabled or accelerated new forms of human interactions through instant messaging, Internet forums, and social networking.”¹

“The Internet can now be accessed almost anywhere by numerous means, especially through mobile Internet devices. ... Within the limitations imposed by small screens and other limited facilities of such pocket-sized devices, services of the Internet, ... may be available.

Service providers may restrict the services offered, and wireless data transmission charges may be significantly higher than other access methods.”²

1 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

2 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

“The low cost and nearly instantaneous sharing of ideas, knowledge, and skills has made collaborative work dramatically easier, with the help of collaborative software. Not only can a group cheaply communicate and share ideas, but the wide reach of the Internet allows such groups to easily form.”³

“World Wide Web browser software ... lets users navigate from one web page to another via hyperlinks embedded in the documents. These documents may also contain any combination of computer data, including graphics, sounds, text, video, multimedia and interactive content including games, office applications and scientific demonstrations. Through keyword-driven Internet research using search engines like Yahoo! and Google, users worldwide have easy, instant access to a vast and diverse amount of online information. Compared to printed encyclopedias and traditional libraries, the World Wide Web has enabled the decentralization of information.”⁴

Future visions? No, these statements are snippets from the most current definition of ‘Internet’ I found — at Wikipedia. Are we ready for using or serving these pictures?

I guess that most of us / most of our institutions have a big fund of data — metadata and even content — in our databases. Are these data open to everyone and open for being linked to other databases as it is requested in that definition of the internet? Do we use information from outside to enrich our own data? Do we foster our efforts to digitize objects that have been analog so far? Some of us do — and others are still very traditional, pointing out that e.g. there is not enough money to implement new processes and services. But moving institutions from the state-of-the-art to new thinking is not only a question of financial means. It is mainly a question of policy.

Policy in the context of digital media and the Internet includes many aspects of argumentation — among them the questions of

- Selection (Which types of objects should be digitized? For what reason? What is meant by ‘quality’ in this context?)
- Access (Open access vs. IPR? Material of poor quality? Material that can be found only once? Different, new access points?)
- Search and retrieve (Standards, networking, automatic indexing?)
- Re-use of data (Licensing? Commercial/non-commercial? Legal framework?)

I would like to give some examples for sensible activities in this context, bearing in mind that decisions often are dependent on formal conditions. Not always, however, not exclusively and not if professional answers have to be given.

Let me start with the selection of objects.

Do we need to digitize everything? Or do we have to keep our capability of valuing the meaning of objects for future generations? According to which criteria? I often discuss these questions with friends and colleagues — questions which are not new, but have been discussed continually over the last decades. When the German National Library started its collection of comics at the beginning of the 1950s, the whole library and scientific world was laughing or mocking about the ‘level of content’ in our magazines. And today? I no longer need to explain this.

Our collection at the German Exile Archive 1933 – 1945 contains documents, letters, photographs, manuscripts, etc. of Germans who had to emigrate during the Second World War. Imagine if we had kept only the personal papers of famous, well-known people! Or if we had thrown away all the material not directly related to the professional works of the persons!

3 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

4 <http://en.wikipedia.org/wiki/Internet> last modified on 12 August 2011

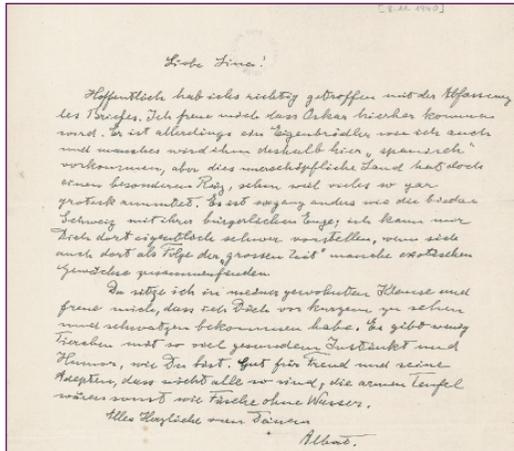


Figure 2: Letter from Albert Einstein to his cousin Lina Kocherthaler of November 8th, 1940

Private letters written by Albert Einstein, allowing for a broader view of his personality, would not exist today.

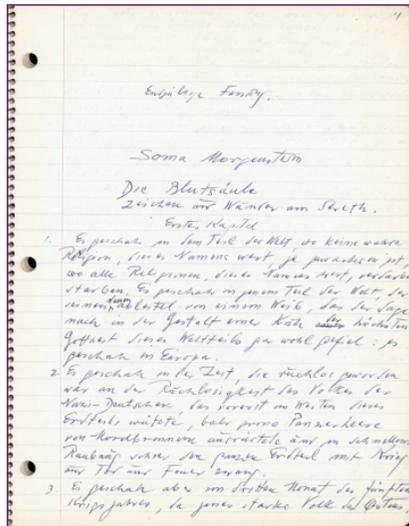


Figure 3: Part of a manuscript of Soma Morgenstern, written between 1933 and 1945, his exposition on the Holocaust, published on the base of the papers at the German Exile Archive during the 1990ies by Ingolf Schulte, Zu Klampen Verlag

The works of Soma Morgenstern would have been lost. Instead they have partly been published again.

I realize that I am speaking of physical media — handwritten or typewritten letters and manuscripts right now. But try to transmit these thoughts to the digital world of today, and think of the possibilities technical developments allow for searching, comparing and working with material that is spread all over the world. And then think of selection criteria and answer to the question: "Do we really know what will be interesting to future generations?"

And, especially when we are looking to our physical holdings, traditional media in our archives, libraries and museums have been collected over the years according to certain rules, collection or selection criteria, in order to fulfill the special tasks of the particular institution.

Digitizing these media means you do not have to do a second selection!

Therefore, looking to our traditional media, digitization is not a question of selection according to certain criteria, but according to financial resources, time line, priorities, quality and/or aims we want to reach. One priority for digitization could be the point-of-view of preserving the originals due to their poor condition. Another one could be the fact that there is only one existing specimen of a certain object. In both cases digitization is essential for usage: either because the original cannot be used any longer, or the original can be found only once — the digital version is accessible at any place and any time. I will come to this again later.

In this context the next decision to make is for the quality of digital objects. We should try to digitize as well as possible, but what does that mean? Should we produce a new product from the old material keeping only the content itself without the accompanying aspects, such as disturbing noises within sound files, little breakouts in films, dark paper in text documents, etc.? It is much more authentic and original to keep all these things together with the digitized object, in order to give users an idea of the original media. Alternatively you decide to create a new product, using the possibilities of modern digitization technology.

My own opinion on this is:

You should digitize as originally as possible, in order to preserve authenticity! Creating new products is on the producers.

I now come to a second aspect of digitization: the aspect of access.

As I mentioned already, through digitization activities, objects often are accessible again that could not be used before because of their poor physical, fragile conditions. The digitization itself of this material is a means both of preserving the originals (I mentioned this already as an aspect of selection) and offering the possibility to work with the content again.



Figure 4 and 5: Examples for the digitization of old piano cylinders

Perhaps you have already heard of our project at the Deutsche National Bibliothek (German National Library) to digitize the old piano cylinders from the end of the 19th/beginning of the 20th century. Some of them were made of wax and could not be given to users at all. A similar situation can be seen with piano rolls — made of paper and being used only with very special pianos, they cannot be given to users either. Three years ago the Music Archive of the German National Library digitized some copyright-free pieces of music that could be found on piano rolls as well as on shellac disc. Now access to this material is possible again for users.



Figure 6: CD, published by the German National Library, German Music Archive, in 2009

At this point I would like to emphasize that I am aware of the fact that access to these objects primarily is dependent upon copyright regulations. But supposing we could deal with and digitize copyright-free material, what is the best decision concerning access to this material?

1. Free access to everybody interested in the objects, or
2. Access restricted to defined user groups e.g. by financial, professional or even local criteria?

Even if digital objects are copyright-protected and therefore access via the Internet normally not allowed, we can offer a lot of information on digital objects via our databases.

Institutions often fear that free access to their digital holdings or even to the metadata on them might make them unnecessary and that people will no longer come to the reading rooms, sound archives, etc. I am convinced that the opposite is the case — using an institution via the Internet only means another type of user and usage. We cannot operate only locally, but regionally, nationally and internationally as well. This could cause the number of users to grow greatly — and these figures can be used for political argumentation and for generating political awareness.

But as a basic requirement for that we **must** provide free access to our catalogues and metadata.

In detail the metadata information includes:

- description of the content/provenance, etc. of the objects,
- description of the technical conditions of the object,
- description of the legal conditions under which the object can be used.

(For the rest of my speech I will concentrate on the first group of metadata.)

These metadata are essential because they are the connection between users on the one hand and objects on the other. The more information you can offer on the subject — on concrete content details, on the context of digital objects — the more people gain knowledge of the digital material in your institution. The wider you spread this information on your digital holdings, the more people like to have access to them, sometimes regardless of whether free or restricted access is possible. That means that your institution does not have to be the **one** single access point, but there should be others via related institutions, search engines and networks.



When preparing for this talk, I was asked whether librarians, archivists, curators of museums and other colleagues do not lose “control” of their digital holdings when giving free access to objects and data. And I was asked how to make sure that digital content and digital data are used in the right context. What do you think? Can we avoid usage of digital information that is different from our own professional or scientific thinking? Do we **have** to avoid this? I can imagine several forms of abuse of digital information from my point-of-view, but often others have different opinions.

And is the use of material in a different manner than usual not often the starting of new scientific discussions? How do you control such processes? No chance!

Access to digital content and data should be as open as possible in the light of copyright regulations and intellectual property rights. If not the objects themselves, the metadata should be freely accessible to make the content attractive to potential users.
Abuse cannot be prevented. Maybe we have some influence on the streams of data on the Internet to a certain extent, but no more.

Offering free access to metadata or even to digital objects themselves has also a lot to do with the question of “how best to prepare the search and retrieval of digital objects?” Traditionally search and retrieve is one of the main questions of our professions. We drafted cataloging or indexing rules over centuries, trying to make catalogues or finding aids as generally usable as possible. The description of objects was aimed at all necessary information our users needed to give them an idea of the single object or of the collection of which the object was a part. We still do this nowadays for analog material, and some institutions try to do the same for digital collections — but does that make sense?

For digitized material it is useful to use the metadata describing the originals. But the mass of born digital material is difficult to index or catalogue the same way. And even if you dare to be selective, it is much more material than ever before.

Additionally, you do not need to describe digital objects as you have done before. Users do not need to get an idea of the object, because, ideally, they can have direct access to it working in an archive or library, etc., or even via the Internet.

So far in my speech I have talked a lot about open access to our data, about Internet and networking, about users who are no longer local. With all that in mind, I believe that today's cataloging and indexing work consists of

- describing the necessary basic information
- analyzing the whole digital object and extracting automatically each additional piece of information you can get,
- combining this automatically generated information with the databases we already have (authority data, etc.),
- offering additional search entries by this.

We do not need to describe objects to include them in our catalogues or finding aids any longer; we need description to put the objects into subject or time or local contexts and offer as many access points from different points-of-view to it as possible.

Let me show you an example of the linking of data.

Link zu diesem Datensatz	http://d-nb.info/351971710
Titel	The Beatles 1962 - 1966 / Lennon ; McCartney
Komponist(en)	Lennon, John ; McCartney, Paul
Verlag/Firma	Köln : EMI-Electrola
Jahr	[1987]
Umfang	2 Schallpl. in Album : 33 UpM ; 30 cm
Inhalt	Erth: Love me do. Please please me. From me to you. She loves you. I want to hold your hand. All my loving. Can't buy me love. A hard day's night. And I love her. Eight days a week. I feel fine. Ticket to ride. Yesterday. Help! You've got to hide your love away. We can work it out. Day tripper. Drive my car. Norwegian wood. Nowhere man. Michelle. In my life. Girl. Paperback writer. Eleanor Rigby. Yellow submarine.
Interpret(en)	Beatles
Anmerkungen	Interpr.: Beatles. - P 1962 - 1966. - Weitere Ausg.: MC, Best.-Nr. 10 5325 4.
Bestellnummer(n)	Apple 1C 172 10 5307 1
Sachgruppe(n)	10b Popmusik
Leipzig	Signatur: SA 67/04955 Bereitstellung in Leipzig

Figure 7: Cataloguing description of a music publication from the Beatles, mentioning John Lennon as one for the composers

Link zu diesem Datensatz	http://d-nb.info/gnd/118571575
Person	Lennon, John (männlich)
Andere Namen	Lennon, John Lennon, John Winston Reno, Jon Lennon, J. Lennon, John Lennon, John Winston Ono Lennon, ...
Quelle	LCAuth; All mus. guide (Internet)
Lebensdaten	1940-1980
Beruf(e)	Rockmusiker Komponist
Instrumente/Vokalstimme	el-gitar (el-g); gitar (g); harmonica (harm); keyboards (keys); piano (p); vocal (voc)
Land	Großbritannien (GB-GB)
Weitere Angaben	Engl. Pop-Sänger und Songwriter, Mitglied der Beatles; Musiker, Performancekünstler, Filmemacher, Zeichner
Beziehungen zu anderen Personen	Lennon, Julian (Sohn); Lennon, Sean (Sohn); Ono, Yoko (Ehefrau)
Organisation(en)	Beatles Stuart Ocho Band
Autor von / Beteiligt an	260 Publikationen <ol style="list-style-type: none"> 1. Power to the people [Hedvinkombi] Lennon, John. - Köln : EMI Music Germany 2. John Lennon - in my own white (Zwilling) Towns, Colin. - Fockwisch : Provocateur Records ; Köln : Aive, Vertrieb, P.2010 3. John, John - John Lennon Berlin : Gr. Schulbücher, 1981 4. The Beatles Berlin : Gr. Schulbücher, [2002], [Dachw.] 5. Power to the people / CD. The music Köln : EMI Music Germany, P.2010 6. ...
Thema in	74 Publikationen <ol style="list-style-type: none"> 1. Das Neue, so merkwürdig ...

Figure 8: Information on John Lennon in the Authority File for Names of the German National Library

Weblinks [Bearbeiten]

- [Commons: John Lennon](#) – Album mit Bildern und/oder Videos und Audiodateien
- [Wikiquote: John Lennon](#) – Zitate
- [Offizielle Website](#)
- Literatur von und über [John Lennon](#) im Katalog der Deutschen Nationalbibliothek
- [John Lennon in der deutschen und englischen Version der Internet Movie Database](#)

Figure 9: Link from the authority file to the information in Wikipedia

John Lennon

John **Winston** Lennon (geb. John **Winston** Ono Lennon; MSD CY 7. Oktober 1940 in Liverpool, England; † 8. Dezember 1980 in New York, USA) war ein britischer Musiker, Komponist und Autor. Während seines Studiums an der King's College School und danach an der britischen Musikgruppe The Beatles, für die er neben Paul McCartney die meisten Stücke schrieb. Das Komponistenpaarwechsel Lennon/McCartney war eines der produktivsten und erfolgreichsten in der Geschichte der Popmusik.

Gemeinsam mit seiner zweiten Ehefrau Yoko Ono veranstaltete Lennon ab Ende der 1960er Jahre verschiedene Aktionen für den Frieden und widmete sich auch zunehmend künstlerisch neuen Gebieten. Nach der Trennung der Beatles gründete Lennon eine englische Solo-Gruppe mit Allen war John Lennon/Plastic Ono Band (1970) und Imagos (1971).

Multimediale Zitate

- 1. [Lennon](#)
- 1.1 [John und Yoko](#)
- 1.1.1 [John Lennon](#)
- 1.1.2 [The Queenies](#)

Figure 10: In the annexes of the information in Wikipedia you find a re-linking to the catalogue of the German National Library, among them Figure 7

Normdaten: [PND: 118571575](#) | [LCCN: n00017868](#) | [VIAF: 195044](#) | [WorldCat](#) | [WP-Personeninfo](#)

Kategorien: [John Lennon](#) | [Gitarrist](#) | [Rocksänger](#) | [Singer-Songwriter](#) | [Britischer Musiker](#) | [Grammy-Preisträger](#) | [The Beatles](#)

Figure 11: Both the authority file itself and the article in Wikipedia are linking to other international databases

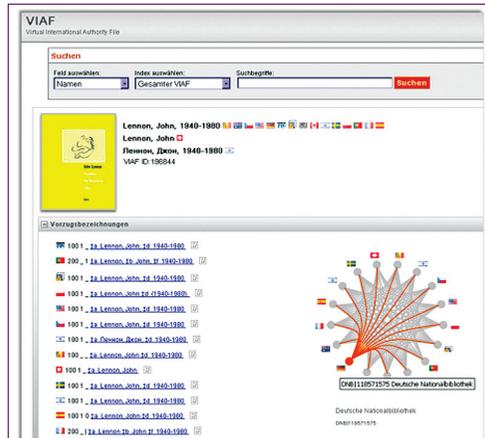


Figure 12: One of the international databases is the virtual international authority file where you then find information on John Lennon from other countries in the world.

In this sense metadata of originals that are digitized could be enriched by extracting additional information out of the content of digital files.

We at the German National Library actually are working on a project, PETRUS, that develops processes for automatic extraction or automatic setting of certain information. CONTENTUS contains extraction aspects for different purposes. Europeana or Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in the context of the German National Library are working on metadata enrichment by automatic processes. I do not want to concentrate on this now, but I'd like to point out that a lot of activities are undertaken in this field and the results can be re-used.

In any case, the metadata of digital objects have to include a persistent identification — that means a machine-readable, trusted address of the digital object that is kept alive “for eternity”. The situation of metadata containing an URL, behind which you see “Error 404 — file not found” is the worst in the digital world and should not happen to institutions like ours.

Summarizing these aspects, I am convinced that sensible digitization also comprises the production of good metadata – intellectual or automatic. “Good metadata” in that sense means useful for our users, not for our catalogues. To retrieve a digital file, you need persistent identifiers.

After having done a lot of work on cataloguing, indexing and possibly enriching metadata, it is of high importance that the metadata can be found by users. For this, another decision in our institutions has to be made: the decision for an open, modern structure of our data formats and for allowing others — even so-called commercial search engines or WIKIPEDIA — to re-use our metadata.

Furthermore, it is necessary for the metadata to have an identifier too. And with this aspect of “re-use of data”, I bring in the idea of linked (open) data.

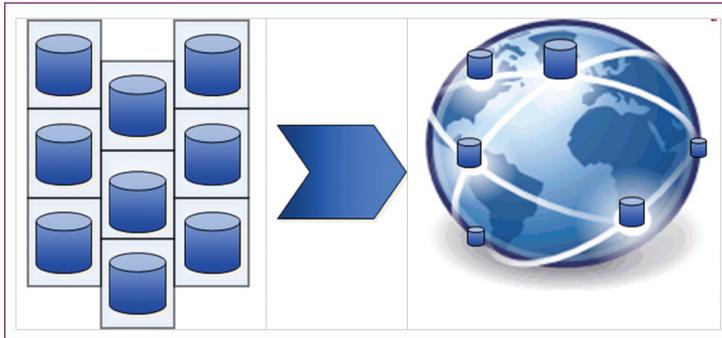


Figure 13: From single data silos to the Giant Global Graph (GGG) — the worldwide database of linked data

The idea of linked data is to move from closed parallel databases of single institutions to one worldwide database that consists of “networking”, of “linked data”.

In 2006 Tim Berners-Lee, the inventor of the Internet, stated: “*The Semantic Web isn’t just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.*”⁵

To realize this, he set up four rules:⁶

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Since then the development of the linked data community between 2007 and 2009 was as follows.

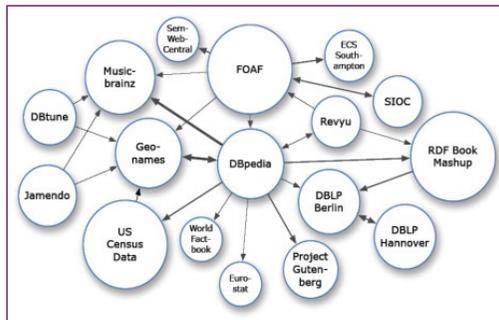


Figure 14: The community in July 2007

5 <http://www.w3.org/DesignIssues/LinkedData.html>

6 <http://www.w3.org/DesignIssues/LinkedData.html>

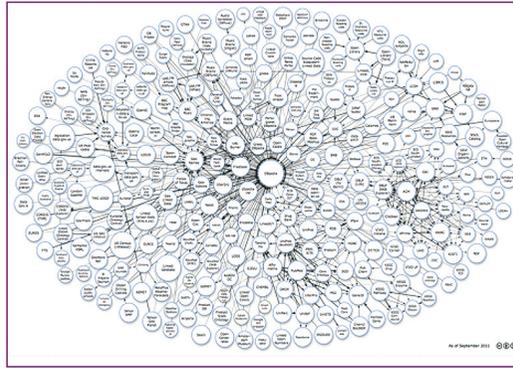


Figure 15: The community in September 2011

I am not the technical expert to explain this in detail, and this is not the right place to do such a precise presentation. But the ideas behind and the advantages of creating direct links between our data are:

- Standards for data publication and exchange already exist on the Web
- The possibilities of reaching more user groups increase (scientific institutions or communities, research institutions and communities, user groups behind other “linked data datasets” which are linked to your own)
- Your own data are enriched by information from other areas, e.g. geographical aspects, encyclopedias, thesauri, biographies, etc.

Practical examples for this can be found at the website of the BBC and at a new museum website in Germany. The BBC has a section about musicians and pop bands on its website, which shows the linking of data very well.⁷ The complete information is gathered from several sources, such as Wikipedia, Dbpedia, Musicbrainz and others.



Figure 16: BBC-Website on the Beatles

For example, if you look for “The Beatles” at the BBC-Website, you find

- the most actual up-to-date information concerning the group or the members of it (selected from the daily information of BBC),
- biographical details (Wikipedia),
- a list of tracks that have been played by the BBC during the past weeks (Musicbrainz),
- a list of the DJs who most currently played Beatles’ songs (Musicbrainz and daily broadcast information)

⁷ <http://www.bbc.org>

- a buyers' guide
- a list of links to other information about the group
- etc.

Another example can be seen with the project “Museum-digital”¹⁸, which is currently in its beta version.

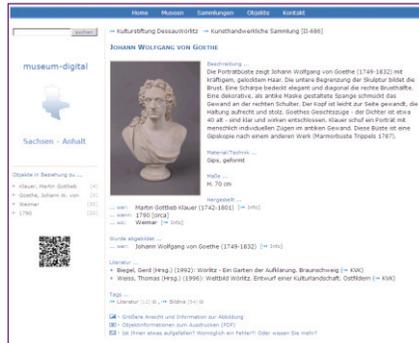


Figure 17: Information on a sculpture showing Goethe on the website of museum-digital

There you find additional information included in the object description on persons who are in any way at all involved with the objects. Or you get geographical details that come out of the “Getty Thesaurus of Geographic Names Online” (TGN). The connection between the description of the objects and the biographical information on persons is done by using the authority file for names of the German National Library.

I think these examples show very clearly the advantages of linking data from all over the world. It is not a linking of the objects themselves, but only of the metadata that describes the objects. Each bit of information contains the link to the objects or to a front page where the users can find a hint of how to achieve access.

This is a big service for users, and therefore I firmly believe that the re-use of data should not be limited to data exchange between certain databases. It should be developed to linked data activities – preferably ‘open’ – to build up the ‘Giant Global Graph’.

Well, my speech is based on my own opinion and the experiences we have had at the German National Library. You might have different point-of-views. But the summary of my recommendations on the sense of digitization is as follows:

Digitization makes sense, if

- as much material as possible is digitized,
- even objects which cannot be used in the original form, because of preservation aspects, are digitized,
- as much material as possible can be accessed via the Internet,
- the information (metadata) on the objects offers as many access points as possible,
- the metadata can be used by search engines and
- also re-used within the linked (open) data community.



Are we ready for that? Do we have the right curricula at our universities and professional schools? Do we address the right issues to our young professionals? Do we have to change the way we see ourselves within the different professions?

And, looking at the framework that is necessary to realize the picture I've drafted: Have we solved the legal or even the technical problems arising from such a picture? I am convinced that librarians, archivists, curators and other information specialists are going into the right direction, but nevertheless there is still a lot to be done.

John Lilly⁹ once stated: "Our only security is our ability to change." Our institutions have to change to play a further role in the changing information society.

The IASA 2011 conference in Frankfurt we just started will be a good platform to take one or more further steps, and I am convinced that we will hear a lot of new activities, learn a lot from the experiences gained from projects and discuss a lot about new developments and strategic directions.

9 John Lilly, http://en.wikipedia.org/wiki/John_C._Lilly

“AND THE SURVEY SAYS...OK, BUT THE DATA SAYS!”

Cassandra Gallegos, *George Blood Audio and Video*

Introduction

Data. The mere mention of the term sends waves of quantitative hope down the spines of librarians. This is evident in the conference programming within the library and archives field. Every year at many library conferences there are multiple presentations on preservation surveys or data collection. The information is presented in its raw natural state, exactly as it came from the survey with little exploration into what the numbers may mean for not only their collection, but for the field as a whole. The data is often silenced before it has a chance to tell us anything of real consequence. Even worse, a librarian runs the risk of becoming incapacitated by the sheer quantity of data they receive. Upon contemplating this data dilemma, I wondered if there is another way to listen to our data and hear what it has to say, because if we do not our survey data can become nothing more than noise.

The preservation field's love affair with data causes unsure librarians to rely heavily on data collected from surveys. They are left to hope that this data will guide them towards making a definitive decision on any number of preservation concerns. What should I digitize first? How much of my collection has problems with mold? Is this object going to deteriorate soon? Unfortunately, data alone cannot tell you what to do and it is unlikely to tell you what you do not already instinctively know about your collection materials. This survey zeal hits fever pitch when it comes to audio and moving image materials. The idea that more data equals a better-protected collection is prevalent even amongst those of us who should know better! But surveys and data are not the enemy. In fact they can be quite helpful if we know what questions we want to ask of our data. The goal of my research is to find new ways of analyzing traditional American survey data along with an analysis of the data collected where I work.

Institutional collaboration

I am the Preservation Administrator for George Blood Audio and Video. We are an audio and moving image digitization vendor located in Philadelphia, PA. Our data collection was initially motivated by a need to provide such metadata to clients following a digitization project; however, I felt that there must be more we could do with this data. In preparing my analysis I worked with additional data sets from Columbia University (New York, NY), Harvard University (Cambridge, MA) and Indiana University (Bloomington, IN), each of which has created a unique survey tool.

Columbia University's data was collected using AVDb, a popular survey instrument created in 2008 by Columbia University with a generous grant from The Andrew W. Mellon Foundation. It is designed to aid in setting preservation priorities for unique and rare audio and moving-image materials. Surveyors are not expected to be experts in audio-visual media, but they are required to have familiarity with broad categories such as format identification. The accompanying users manual provides sufficient support to fill in any knowledge gaps. Applicable either to an item-by-item inventory or a random-sample survey, AVDb does not require collection of a great deal of detail about each item, yet has the ability to be a very granular tool. Data is collected purely through visual inspection. No playback of any medium is required.



<http://library.columbia.edu/services/preservation>

Harvard University created a similar tool called S.A.V.E. (Support for Audio-Visual Evaluation). This web-based tool enables the collection and aggregation of item-level condition assessment data for audio-visual materials that unites physical characteristics and condition data with curatorial information related to an item's research value. The survey taxonomy was designed to support the multi-use and item-level physical condition assessment survey of all audio-visual materials across the Harvard University campus.



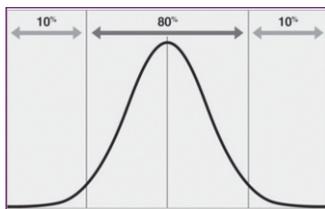
<http://preserve.harvard.edu/index.html>

Indiana University Bloomington Campus created FACET (Field Audio Collection Evaluation Tool) to survey the various audio-visual collections on campus. FACET was created by Michael Casey in the Archives of Traditional Music. It is a point-based, open-source software tool that ranks audio field collections based on preservation condition, including the level of deterioration they exhibit and the degree of risk they carry. It was designed to assess the characteristics, preservation problems, and modes of deterioration of the audio formats historically associated with field recordings rather than those formats developed for commercial recordings. Similar to AVDb, FACET has an accompanying user manual to aid in the identification and assessment process.



http://research.iub.edu/communications/media_preservation/

Columbia, Harvard and Indiana Universities' survey tools have collected rather large quantities of data. The Law of Large Numbers tells us that large amounts of collected data have a common distribution pattern and that most observations hover around the average; the odds of a deviation declining faster and faster as you move away from the average, resulting in a standard Gaussian distribution. In other words, the more data you have the more likely it will assume a bell curve.



Data analyzing procedures

As a vendor, we are in a position to harvest the metadata we collect from a wider range of institutional situations rather than the data from any single institution such as is the case with university data. Additionally the difference between the data collected with AVDb for example and the data collected at our facility is that our data is post-digitization, so we can know much of these data fields with certainty. Survey data often includes a lot of guesswork. For example, an archivist must rely on notes from reel boxes and cassette shells for length, speed, track formation and other various important data fields. Often these notes are wrong or the box is mislabeled which in turn misleads the archivist and skews the survey data.

At the time of my initial data collection, metadata on 33,000 objects was available in our database. These factors, large number of institutions, wide range of collections, large number of data points about *actual* samples, allow me to consider our data set as representative of the general population.

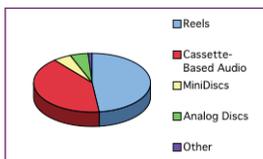
I have used this general population, the George Blood Audio and Video production data, to spot trends, then make hypotheses based on these trends. I tested these hypotheses by examining the similar data points from the Columbia, Harvard and Indiana surveys. In what ways do the surveys gather information that is similar to or different from the production data that we collect? Furthermore, I wondered if it was possible to extrapolate for unknown or uncertain data in surveys based on known information from our production data to help librarians and archivists make more informed guesses about their collection materials.

In analyzing the data I needed to control for a few factors. Some institutions collect their data on a different basis than others. For instance, does the “intellectual unit” or what I would consider a single record in a database, correspond to a single physical object or a single interview. Another challenge is that new records are constantly being added to our production database. For this reason I used 1%, or about 300 objects (about one week of our production capacity), as the margin of error in my analysis.

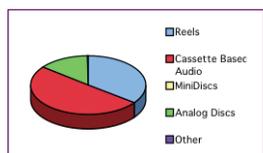
Additionally, each survey collected information in slightly different ways and I had various levels of access to each data set. For instance, Columbia’s survey groups audiocassettes, microcassettes and DATs in one category. Therefore I had to do the same grouping with the other three surveys to make the data directly comparable. Columbia and Harvard collected data at the object level while Indiana collected information on collection level by format. Both Columbia and Indiana provided me with the actual database and data used during the survey. Harvard’s system could only be accessed in-house, so I submitted queries directly to Harvard via email to obtain my results.

Results

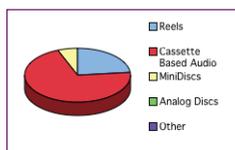
The first data I collected concerned format and length of audio. At George Blood Audio and Video we assessed 32,699 audio objects at the time of my data collection. Out of this total, 15,390 are reels and 13,399 are cassette-based audio formats. This accounts for nearly 90% of the audio we handle. After these two formats, numbers take a drastic downturn. 5% are MiniDiscs and 5% are analog discs. Cylinders, Wires, and CDs account for less than 1% each.



Similarly, Columbia’s surveyed collections are 85% magnetic tape, either reels or cassette-based, with approximately 35% being reels and 50% cassettes out of 35,766 total audio objects.

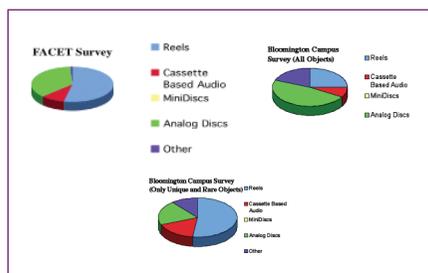


Harvard audio survey of 3,839 objects is 69% cassette and 22% open reel.



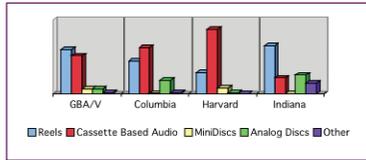
Indiana University's survey has a total of 364,867 audio objects across the entire Bloomington campus. FACET was used to survey objects within one collection; however, another campus-wide survey looks at the number of formats across all collections. I parsed both sets of data and came to the following conclusions: The FACET data shows that this one surveyed collection is approximately 53% reels, 36% analog discs and nearly 10% cassette-based formats. Up to this point in my research, it had become clear that the average collection consisted of mostly reels and cassette based media. Indiana consists of mostly reels and analog discs. I wanted to delve further into the Indiana data because I noticed that Indiana made a distinction between objects that were unique and rare and those that were commercially available. Unique and rare objects are more of a preservation concern than those that are commercially available. Unfortunately this information was not quantifiably available within FACET. There were various content notes and a point system dedicated towards the uniqueness of each subset of the collection. This data was not useful without intimate knowledge of the initial data collection process. I turned toward the inventory report for the entire Bloomington campus to shed some light on whether Indiana's collections consisted of mostly reels and analog discs that were unique/rare or if the inclusion of commercially available objects were skewing the data. This report extrapolated the type of data I was unable to process within FACET. The initial results within this report indicate that overall the collections on the Indian Bloomington campus consist of a large percentage of discs (47% analog discs) much like the FACET survey.

Upon further inspection of the objects indicated as either unique or rare, only 15% of the total collection is unique or rare discs that are not available commercially, 21% of the total audio objects are unique or rare reels that are not available commercially and 5% of the total audio objects are unique or rare cassettes. These results align more with my hypothesis that the average is mostly reels and cassettes.



It is noticeable that 21% of the total is devoted to various other audio formats. Taken as a whole this 21% appears to be a substantial part of the collection but within this 21%, 82% are optical discs only of which 14% are unique or rare and not commercially available. When you adjust the data to only include the unique or rare non-commercially available audio objects, you see that the total amount is really 6% unique and rare objects of various other formats.

With all of the data together it is clear that most of each collection is either reels or cassettes.



As far as reel preservation is concerned, I know anecdotally, from handling material, that many more reels suffer from poor tape pack rather than biological contamination or chemical degradation. I tested my experience against the data collected at Columbia. They assessed that 90.59% had a poor library wind. This information is critical for an archive thinking about creating an in-house audio digitization department: it may make sense to have a large-scale library wind project.

After spotting these specific trends in collection content, I considered what further questions could I ask of the data. Librarians and archivists in general are concerned with issues of degradation such as sticky shed syndrome and contaminants such as mold. I often receive questions on baking reels from archivists — how to do it, when to do it, and if it should be done at all. I would not be surprised if there are portions of collections out there that were accidentally destroyed by a well meaning archivist.

I wanted to know, from our production data, exactly how often stickiness occurs in reels. Surprisingly I cannot tell you! At our studio it has been our experience that most reels that arrive exhibiting signs of sticky shed syndrome will “get better” after spending 6-8 weeks in our temperature and humidity controlled vault.¹⁰ The time a tape needs in the proper storage environment to diminish sticky shed varies based upon the width of the tape. We have experienced that a reel must remain in our vault for 6-8 weeks per each ¼ inch in width. This data is merely anecdotal and will require further study. The reels that are still sticky after their stay in our vault are reels that require baking. While I cannot tell you how often sticky shed syndrome occurs within the general population, I can tell you, within the general population, how often reels will still require baking even after stored in a proper, low humidity environment.

Around 6% of all the reels we received required baking. And this is highly concentrated within collections. If kept in the correct conditions for longer, who is to say this may not “cure” more of the sticky reels. Specifically it seems that tightly controlled relative humidity will help these very sticky reels. The data also tells us that sticky shed is not a large problem within the general population but rather a concentrated problem within collections. If you find one sticky reel in a collection, there is a higher likelihood that you will find more sticky reels within the collection. This is something that a random survey may miss. Furthermore, proper storage conditions should be the primary goal for several reasons. First, there is likely only a small chance of encountering a bad case of sticky shed syndrome. Second, since this chance is indeed so small it is more cost effective and safer to send these few reels out to a professional rather than risk these materials by baking the reels yourself.

As for mold, less than 1% of reels in our production data had mold. Mold occurs only rarely on other media. Jim Riley at the Image Permanence Institute speculates this is due to the nature of the exposure of the media. Magnetic media does not contain anything that mold will live on. Mold is growing on the biofilm that develops on objects when exposed to the environment and human handling. Since reels are more exposed, more biofilm develops and therefore supports the growth of mold. So while mold and sticky shed syndrome should be taken seriously, they do not occur very often in the general population. Worry and effort should be placed elsewhere. Be aware of these issues but only so much as to know to seek information and help when these problems arise.

¹⁰ The set points are 65°F and 35% R.H.



I asked further questions of my data concerning reels. I wanted to know how many of each stock brand required baking. Is there a correlation here? First off, the two tape stocks we encounter most often are Scotch and Ampex, with Scotch coming out as the clear leader; however, we only baked 2% of the Scotch while we baked 19% of the Ampex.

Amongst compact audiocassettes I found that both Columbia's and the George Blood Audio's cassettes were approximately 92% Type I, which is not overwhelmingly surprising as they were the most affordable type of cassette. What is surprising is that 50% of the cassettes we received needed their record tabs removed. Removing record tabs is the cheapest and easiest precaution you can take towards safeguarding a cassette collection, yet half of all cassettes were left without write protection. In some cases, whole collections were in danger while other collections only had one or two cassettes with record tabs not removed. Additionally we know that replacing the shell solves most cassette playback problems. We replaced cassette shells on about 3% of all cassettes.

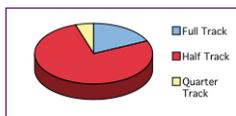
Although we do not see much mold on cassettes, I did notice that, while assessing a large batch of thousands of cassettes from a collection from Puerto Rico, that there seemed to be a correlation between mold and the opacity of the cassette shell. I pulled all the data from all of the cassette assessments done at George Blood Audio and Video. Over 90% of the moldy cassettes were from this Puerto Rican project, so I decided to concentrate on these 1,385 cassettes. Puerto Rico is an ideal location for mold growth with high temperature and high humidity. Out of all of these cassettes, 804 are clear and 581 are opaque; 92% of the moldy cassettes had opaque shells. I also observed that the mold on the tapes housed in the opaque shells was not present in the small clear window that exposed the tape to sunlight. Despite this data you must remember that correlation does not necessarily indicate causation. Further scientific study is needed to prove that cassettes with opaque shells are more likely to be moldy.

Surveys can only collect known information. Object duration is one important datum that is often not known at the time of the survey. Many surveys allow you to collect "maximum duration" based on the possible duration of the audio object. For instance, you assume a C-90 cassette will be full of 90 minutes of audio. Our production data can answer the questions of average and distribution of durations of audio objects. Here are the average durations of various audio formats:

- Audiocassette – 53 minutes
- Reel – 50 minutes
- MiniDisc – 60 minutes
- Analog Disc – 6 minutes
- Microcassette – 30 minutes
- DAT – 69 minutes
- Cylinder – 2.4 minutes
- Wire – 43 minutes

Track configuration of reels is also not often known until time of transfer, unless the archivist owns a magnetic viewer and has an inordinate amount of time to discover the track configurations of all of the reels within the collection. Even with a magnetic viewer, it is generally not possible to discern two track mono from two track stereo simply by looking. I found that a reel will be full track around 18% of the time and quarter track about 5% of the time. The remaining 77 % are half track.

The distribution of speeds among reels varied greatly:



7 1/2 ips (inches per second): 45%

15/16 ips: 2%

1 7/8 ips: 15%

3 3/4 ips: 20%

15 ips: 16%

30 ips: 2%

Cassettes did not vary as much in speed, 99% were 1 7/8 ips.

Conclusion

Surveys tools can be quite helpful if you know how to use them and do not rely too heavily on the data to form conclusions. A more accurate snapshot of a typical collection can be gained by looking at production data that comes from a digitization project. What I have learned is that there is a good chance that the typical American audio collection will consist of cassette-based and reel-to-reel tapes. The reels are not in great danger from sticky shed as long as temperature, and more importantly relative humidity are controlled. If sticky shed is found during a random sample survey then it is likely many other reels within the collection will also be sticky. Mold is a relatively small problem for reels and even smaller for cassettes but there may be a correlation between cassette shell opacity and mold. Additional data points and sample collections will shed more light on these and other issues in audio preservation.

Acknowledgements

I would like to thank individuals and their institutions for sharing their survey data and help preparing this report:

Janet Gertz, Director for Preservation, Columbia University Libraries; Elizabeth Walters, Program Officer for Audiovisual Materials, Weissman Preservation Center, Harvard University; Mike Casey, Associate Director for Recording Services, Archive of Traditional Music, Indiana University; George Blood; Annie Peterson; Melissa Hernandez; Michael Mullin.

SIMPLIFYING SEMANTICS: LINKED DATA RULES OK!

Fran Alexander, *BBC Information and Archives*

1. Introduction

This paper is based on a presentation given to the Organizing Knowledge Taskforce at the 42nd IASA annual conference in Frankfurt in September 2011. It gives an overview of some of the ideas and principles that underpin Linked Data, Open Data, and the Semantic Web. It offers some case study examples and suggests some ways to start thinking about Linked and Open Data projects.

2. Evolution of Knowledge Organisation Systems (KOSs)

Metadata is nothing new. Fashions for declaring the author and title of a work have changed over the ages, but for centuries archivists, librarians, cataloguers, and indexers have been working out how to make sure they know what content and assets they hold and how to make those works or artefacts findable. We catalogue, we describe, we classify. The Semantic Web and Linked Data take those principles into the world of computers and automated processing.

Very early on in the history of human literacy, the usefulness of the **label** became apparent. If you fail to add labels to things, you lose them, or you forget what they are and what they mean. Then people realised labels could be grouped into **lists** (and eventually controlled vocabularies, keyword lists, tag lists, and folksonomies). Some of the oldest forms of writing known to history are lists and catalogues from ancient Sumeria.

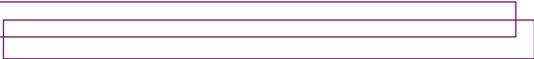
Once lists started to get long and unwieldy, people broke them up into sections or categories and formed the first **taxonomies**. Taxonomies were known to the ancient Greeks, although it was the biological taxonomies of Carl Linnaeus (1707-78) that made them a key scientific tool. Taxonomies are easy for humans to understand, especially to provide vertical “drill down” navigation. This means they are used for folder structures, website navigation, and for applying tags. Related taxonomies can be joined together for richer information structuring as **faceted** or **polyhierarchical taxonomies**, labelling different aspects of a concept with different facets, which are useful for refining search results.

More recently it was noticed that if you specified and defined the relationships between the facets (or terms and concepts) as **ontologies**, you could use computers to perform very sophisticated processing of complex queries, including processing Linked Data. Ontologies are often shown as diagrams that look more like topic maps with lots of links and connections rather than a simple tree diagrams.

3. What is Linked Data?

Archivists, librarians, and knowledge managers already work with computerised systems and already use “linked data”. Cross references are links. Multiple types of indexes enable us to find resources in different ways, by looking up from different starting points (by title, by author, by date of purchase). There already exist standard formats for creating, coding, and publishing data so that it can be shared, either within organisations or with the wider community, for example MARC (MACHINE-Readable Cataloging) or Encoded Archival Description (EAD) — an XML format. Linked Data is another way of making records machine readable and interoperable. Linked Data uses a format called RDF (Resource Description Framework) for publishing or exporting data. Linked Data can be shared easily because it follows several key principles:

- **Use identifiers:** ISBNs identify specific books and primary keys in databases identify particular records. Linked Data identifiers are known as URIs — Uniform Resource Identifiers. A familiar form of URI is the URL — Uniform Resource Locator — used for website addresses.



By using public identifiers, for example to identify concepts in a taxonomy, instead of words, it is clear what you are referring to. An example would be the disambiguation of jaguar the big cat from jaguar the type of car. More significantly, computers can automatically associate every time you have used the concept jaguar (the cat) with uses of jaguar (the car) in somebody else's data, because the URIs will match, so avoiding the problems of understanding ambiguous words or resolving differences in spelling.

- **Provide descriptions:** When you publish your data, you should explain what your data means. This usually takes the form of public identifiers and an ontology showing how the concepts you have used relate to each other. If you use URIs as identifiers, you can provide extra information about the concept yourself. If you use a published open URI, you simply re-use the work that others have done.
- **Include links:** Providing hyperlinks to your sources, to other similar data, and to other sites that are using your data, and any other resources that might be interesting or useful will also help others to put your data in context, understand it, trust it, and use it.

4. What is Open Data?

Open Data is data that is available on the open web for others to reuse. It should be published under an open licence (and it helps to make sure that the licence terms can be found easily). The Creative Commons licence is an example of a flexible, open licence. Open Data also needs to be in an open format, so that others can reuse it without having to buy specialised software in order to access the data.

Linked Data only becomes Open Data when you make it available publicly. It is possible to use Linked Data to share data entirely within the firewall of an organisation. Open Data is only Linked Data if it is in Linked Data formats. This is why people talk about Linked Open Data, or Linked and Open Data.

5. Case studies

A growing number of organizations around the world are publishing data in Linked Open Data formats. There is a lot of activity in the academic life sciences community and a large number of libraries are using Linked Open Data formats to publish their bibliographic and cataloguing data. Libraries have a long tradition of sharing cataloguing records but Linked Open Data can provide opportunities to do more interesting things with data than simply merging records databases.

The teams at the Dutch Parliament (Tweede Kamer) realised that their data is ideal for a Linked Data approach, because parliamentary procedures are highly defined and well structured. In collecting the data, they gather a lot of associated information. For example, of every word spoken in parliament they know when it was spoken, by whom, on behalf of which party, in which debate, etc. One of the most popular ways they have visualised the data is into “ataquograms” — colourful diagrams illustrating which MPs interrupted others the most during debates and which MPs were interrupted the most. The Guardian newspaper in the UK used a similar process to produce “League Tables” of MPs’ performance, voting, and attendance records.

to pay editors to write articles about every animal, and so syndicating the free content that Wikipedia and other sources provide is a cost-effective way of giving users extra information. The website carries a disclaimer so that readers know that the content is from Wikipedia, not the BBC:



“This entry is from [Wikipedia](#), the user-contributed encyclopedia. If you find the content in the ‘About’ section factually incorrect, defamatory or highly offensive you can edit this article at Wikipedia. For more information on our use of Wikipedia please read our FAQ.”

6. Challenges

Linked Data and Open Data are not magic bullets. They cannot solve all your information management problems, but they are useful tools for achieving particular goals. The key to using them effectively is to devise a project with a clear scope and purpose that suits a Linked Open Data approach.

Linked and Open Data standards were devised for publishing, not for archiving, preservation, or long-term access. This means that you should consider how to preserve core archive data as an entirely separate issue. In general, the simplest most technology-agnostic formats tend to be the best for long-term preservation purposes. However, making sure that you can export or publish core archival metadata in Linked and Open formats means that you can gain the benefits of joining the Linked Open Data community, while preserving your underlying data sets.

Open Data needs to be data that is free from rights, licensing, data protection, and other legal constraints. However, by thinking through the sources of data that you have, you can usually find some data that you are free to release. Rights and licensing may cause problems for content, but often not for metadata. So, you may be able to publish your catalogue, even if you cannot easily make the assets themselves available. You could not publish the names and addresses of your visitors, but you could publish anonymized data about numbers of visitors or most frequently accessed collections. If you start to think about your metadata as a form of content in itself, you may find you have much data that you use operationally and are free to release. You may not be able to think of a use for it, but if you release it into the Open Web there is a chance somebody else will be interested.

Linked and Open Data and Semantic technologies alone will not solve your metadata curation and governance problems. They may in practice make such issues worse. If there are flaws and inconsistencies in your underlying data, publishing the data will reveal those flaws. For example, if you started collecting date information in US-style month then day, then a few years later switched to European-style day then month, then a few years later to a fully numerical format, your date information will be inconsistent and automated processing of dates will generate errors.

Linked Data depends on using identifiers so that data can be mixed and matched, but making sure that those identifiers are kept up to date and consistent requires some curation and governance effort. If you want to merge or mash up your data with someone else's you will need to check that they have used terminology in the same way that you have. For example, suppose you have some data on the number of international visitors to archives and you find a data set about the total number of tourists who visit European



countries, you could make an interesting comparison to show whether there is a correlation between total numbers of tourists and visits to archives. If both data sets include “Northern Europe”, before you identify your concept of “Northern Europe” as the same as “Northern Europe” in the other data set you should check that both have included the same countries or parts of countries under “Northern Europe”. If the definitions do not match, the figures will not be comparable and mashing up the data will generate nonsense. So, you cannot assert that your identifier for Northern Europe is the same as their identifier for Northern Europe unless you have checked that they really do refer to the same things. If you subsequently decide to include a different set of countries in “Northern Europe” in a new version of your data, you should also adopt a new identifier as you have effectively created a new “Northern Europe”. This sounds complex, but it is a natural continuation of the problems of maintaining consistency in indexing and vocabulary control that cataloguers have been grappling with for centuries.

Another challenge is finding technical staff to help you put your project into practice. However, the technical issues tend to be easier to resolve if you have a clear idea of what you would like to achieve with your data. Existing technical staff may have enough knowledge already or be willing to find out what they need to do and there are many free tools and sources of support within the Linked and Open Data community. There are consultancy firms which can help with specific projects, and individual freelancers or consultants are another option, but it is still useful if you start with an idea of what you would like to achieve. Finally, if you have no funds, publishing your data in whatever format you can and putting out a call for help is worth a try!

A Linked Open Data project is a publishing project, and so needs to be considered and assessed as such, and as with any project the key to success is to define a clear scope and aims. A small, well-specified pilot project is often a good way to start.

7. Opportunities

Although Linked and Open data projects require some thought and effort, they provide interesting opportunities without the need to spend lots of money on software or infrastructure. Even if you only want to release your data in a constrained way, to share amongst a group of universities, or even a group of departments within your organisation, Linked Data principles may help break down content silos.

By openly publishing your data, you promote your organisation and use and re-use of your assets. In the commercial world, catalogues are given away free because they are seen as marketing. Shopkeepers do not consider charging for their catalogues, because their catalogues and metadata are the way they draw customers into their shops, let people know what is available, and advertise and market their wares. Libraries and archives are beginning to see how releasing their cataloguing metadata can work in just the same way, showing people what they can find and encouraging them to engage. When that metadata is published in Linked and Open formats, networks of catalogues can be built up and associated so that when someone searches for something in someone else’s catalogue, they can be given a link to related things in your catalogue. By co-operating in this way you provide a richer service to researchers, some of whom may not have known to look for you directly. In the past, when you had to physically enter an archive or library to access content, there was perhaps less appeal in reaching out to a non-local audience, but online access puts you in reach of anyone anywhere in the world. Now, a researcher in Poland might be delighted to find their local archive’s data linked to that of an archive in New York or Sao Paulo or Sydney.

As well as making your data available for others to use, it is worth thinking about how using the Linked Open Data sources that are freely available may help you. One example might be that you would like to add biographies of authors to your catalogue. Instead of commissioning writers and editors to research and write a set of biographies for you, you could look for a Linked Open Data source.

Other benefits of publishing your Linked Data openly are that you can use comparison with other data sets as a way of discovering buried inconsistencies and quality problems in records and benefit from free “crowd sourced” help in spotting and correcting errors.

Crowd sourcing is a recent term for another familiar idea — accepting comments and corrections from your readers, visitors, and other members of the public. There have always been people who have written to archives, libraries, publishers, and museums to point out errors and omissions or make suggestions and volunteers who have offered to work in the archives for their own interest and desire to participate in something culturally valuable. In the past, they had to do this by working in the archive in person or by writing letters, which required quite a high level of commitment. Online, it is very easy to fill in a form or send an email, so it is easier to get people to contribute. Online, anyone who has access to a computer is a potential visitor, so this increases the number of people who are likely to find and use your data and it is easier to gather communities of specialists, as you are not restricted to only those people who can visit in person. In addition, people enjoy contributing and if you can find ways of making contributing fun — such as by turning improving your data into a game — you are likely to attract more helpers. An example of such a project is the Transcribe Bentham project at University College London (<http://www.ucl.ac.uk/transcribe-bentham/>) where volunteers type and proofread sections of Bentham’s handwritten letters.

Finally, it is worth understanding how new techniques and technologies, such as semantic technologies work, so that you can decide whether or not they would be useful or helpful in other business processes. For example, as speech-to-text processing software improves, it may provide a way of generating transcripts of audio content that has not been cost-effective in the past. Such techniques will result in large text repositories that need indexing and linking, and automated semantic techniques may be the easiest way to do this.

8. Some questions to start you thinking

In order to set up a successful project you need to think through what the challenges and opportunities mean with regard to your particular organisation and circumstances and then to begin thinking creatively about the data you gather and store as an asset in itself.

Here are some suggestions of questions that are intended to help inspire ideas. They are not intended to be a definitive checklist, as no two projects or two organisations are the same, but they highlight general areas that are often worth considering.

- **Who are our key customers?** By thinking about your customers, readers, users, and audiences, you can think about the sort of data that they need and use and what other data they might find useful. Are there questions that are frequently asked? Are there any questions that they ask that you find hard to answer? Would they like more maps, or biographical data, or different kinds of indexes? Is there a particular institution, publication, or source that they use alongside the sources that you provide?
- **How could we present our data in different ways?** Do you provide lots of alphabetical lists? Could those be re-sorted by theme, or chronologically, or vice versa? Could some of your data be visualised on a map or a timeline? How can different types of data be connected?

Many of the techniques that are used when creating exhibitions and special collections apply just as much to metadata as to the assets themselves. Do you have information about birthplaces of authors stored separately from information about events that happened in those places that you could link to place authors in historical context? Could you bring together catalogues of different types of assets and find connections based on indexing terms, or date of acquisition, or how often an artefact has appeared in exhibitions?

- **How can we revitalise legacy content?** Do you have indexes and catalogues or other sources of data that are hardly ever used or referred to? Why is this? Is it because they are hard to access? Are there collections that are hidden because they are not well indexed and so worth opening up to the public to see if anyone actually is interested?

One of the benefits of choosing some little used or neglected legacy content in a pilot project is that it represents a low risk, but potentially big reward if the project is successful.

- **What data is out there?** Simply finding out what is available for reuse, which other organisations are publishing Linked and Open Data, and what projects are being undertaken can serve as a source of inspiration. For small organisations, offering some data to a larger organisation to include in a specific project could be a very cost-effective way of getting technical support and practical help as well as gaining knowledge and skills.

9. Conclusions

The world of Linked and Open Data is one that we have been working in for years. The difference now is that we have new standards and formats and new ways of using our data with the aid of powerful computing. Although there are pitfalls to avoid and obstacles to overcome, if you find out what Linked and Open Data can and cannot do, you can start to think of projects that could benefit your organisation. As information professionals, we should be the ones who best understand our metadata, best understand how to evaluate external sources, and so are best placed to devise innovative and interesting projects to promote our content, our data assets, and our organisations and institutions.

Fran Alexander is Taxonomy Manager, BBC Information and Archives. In 2009 she was awarded a Master of Research degree by University College London, her dissertation proposing a framework for assessing the subjectivity and objectivity of taxonomies, based on original research into 15 major commercial and academic taxonomy and classification projects. She blogs at <http://www.vocabcontrol.com/>.

Some further reading and resources

An Introduction to Linked and Open Data for Information Professionals:

<http://web.fumsi.com/go/article/share/64146>

Linked Data is blooming:

http://www.readwriteweb.com/archives/linked_data_is_blooming_why_you_should_care.php

Nodalities blog: <http://blogs.talis.com/nodalities/>

Linked Data.org: <http://linkeddata.org/>

Are you a semantic romantic? <http://www.vocabcontrol.com/?p=213>

Linked Open Data in Libraries Archives and Museums: <http://lod-lam.net/summit/>

Semantic Web Conference (academic): <http://iswc2011.semanticweb.org/home/>

SemTech (technology focus): <http://semtech2011.semanticweb.com/>

Museums and the Web Conference: <http://conference.archimuse.com/mw2011/about>

Freebase: http://wiki.freebase.com/wiki/What_is_Freebase%3F

MusicBrainz: <http://musicbrainz.org/>

DBpedia (derived from Wikipedia): <http://dbpedia.org/About>

Case Studies

Tweede Kamer (Dutch Parliament): http://www.fed-parl.be/parliaments_net/pdf/_Nelleke-Aders.pdf; <http://staff.science.uva.nl/~marx/pub/adersgielmarx/>

UK government and The National Archives: <http://data.gov.uk/linked-data>

BBC Wildlife Finder: <http://www.bbc.co.uk/nature/animals/>

BBC programmes ontologies:

<http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml>

NoTube – Semantic TV: <http://notube.tv/>

The New York Times: <http://data.nytimes.com/>

The British Library: <http://www.bl.uk/bibliographic/datafree.html>

Civic Apps: <http://www.civicapps.org/>

Walking Through Time: <http://www.walkingthroughtime.co.uk/>

Glossary

API — Application Programming Interface, a “gateway” or interface through which computers can exchange information.

Dereferenceable — you can look it up. A dereferenceable URL can be followed to find or access further data, documents, etc.

Instance — an object of interest — for example a concept that has a URI tag (e.g. lion). The data about instances is set out as triples — e.g. lions eat antelopes. The ways triples can be created is set out in an ontology — e.g. there are lions and there are antelopes and what happens is that lions eat antelopes. Ontologies can be presented as diagrams.

Ontology — a knowledge model or “world view” that a computer uses to process metadata.

OWL — Web Ontology Language — a computer language that ontologies can be written in.

RDF — Resource Description Framework — a computer language that triples can be written in.

SPARQL — a computer language that is used to process RDF to answer questions — e.g. tell me the names of all artists with paintings in the museum painted after 1975.

SPARQL endpoint — a “gateway” or interface to a set of data that accepts queries in SPARQL. A SPARQL endpoint can be public, to allow free and open access to data.

Triple — semantic metadata statement made up of three parts – subject, predicate, object — e.g. lions eat antelopes, Goethe is the author of Faust, Faust is a poem, etc.

Triplestore — a data store designed to hold triples, rather than relational database tables.

Unique identifier — a “key” usually a number or a mixture of numbers and letters that uniquely identify something. URIs are unique identifiers; some are dereferenceable (http URIs) and some are not.

URI — Uniform Resource Identifier – a “key” that identifies a specific concept or relationship. Http URIs can be linked to (they are dereferenceable). URIs can be URLs — Uniform Resource Locators — which we are familiar with as website addresses.

XML — Extensible Markup Language — a computer description format. XML is one of the formats that can include RDF and OWL.

Web services — (in some contexts) “gateways” such as APIs that can be provided so that data can be accessed easily, usually from a website, but sometimes from an internal data store. Web services could be any service provided over the web (e.g. web-based email, cloud computing).

MAP YOUR VOICE: A USER-GENERATED ARCHIVE OF ENGLISH ACCENTS

Richard Ranft, Head of Sound & Vision, The British Library

Paper presented at the IASA 2011 Annual Conference, Frankfurt, Germany

Comparative phonology

Among the 150 million items in the British Library collections are 3.5 million recordings of music, broadcasts, environmental sounds and spoken word recordings. The latter include significant examples of languages, accents and dialects, used in particular by researchers, teachers and actors. Some of the recordings are recitations by different speakers reading out prescribed texts, a practice that enables linguists and phoneticists to easily compare one recording with another. One of the earliest standardised texts used in this way is the *Parable of the Prodigal Son* from the New Testament gospels. This biblical passage forms the basis of part of a collection of recordings (of which the British Library holds a copy) deposited in the Berliner Lautarkiv at Humboldt University, of British prisoners of war in captivity speaking in their natural accents and dialects¹¹. The recordings, made between 1915 and 1918, are possibly the oldest linguistic utterances by ordinary speakers to be preserved. The *Prodigal Son* text was also used in another collection held at the Library, the 1913-29 *Linguistic Survey of India*. The academic value of this particular text derives from the fact that it permits a comparative analysis of several grammatical features such as personal pronouns, noun declensions and present, past and future tenses.

The British Library has been making available some of these dialect and accent recordings online for research and educational uses. Our *Sounds Familiar?* accents and dialects map¹² plots the geographic location of recordings that reveal variations in the pronunciation and vocabulary of speakers around the United Kingdom. Schools in Britain are invited to add their own contributions to the project. The *Sounds Familiar?* contributions use a different standardised text, a passage from *Mr. Tickle*, published in 1971 by Roger Hargreaves (1935-1988), an English author of children's books. *Mr. Tickle*, the first book in the *Mr. Men* series, was written when the author's six-year old son asked "What does a tickle look like?"

*"It was a warm, sunny morning. In his small house at the other side of the wood Mr Tickle was asleep.
You didn't know there was such a thing as a Tickle, did you?
Well, there is! Tickle's are small and round and have arms that stretch and stretch and stretch. Extraordinary long arms!
Mr Tickle was fast asleep. He was having a dream..."*

The *Mr. Tickle* extract takes about five minutes to read. It has words and phrases that are easily recited by any speaker of English including young and non-native speakers and it encourages a relaxed style of delivery. Its academic interest lies in the fact that it happens to encompass most of the 24 lexical sets used in spoken English as identified by phoneticist John Wells (Wells, 1982). Each set is identified by a single keyword that refers to groups of words with the same vowel sound. For example, pronunciation of the word *bath* has a clear north-south divide in England, being pronounced with a long [a:] as in *cart* in southern England, while in the north of the country it is spoken with a short [æ] as in *cat* (Robinson, 2010). Some other examples of lexical sets are given in Figure 1.

11 Available online at: <http://sounds.bl.uk/Browse.aspx?category=Accents-and-dialects&collection=Berliner-Lautarkiv-British-and-Commonwealth-recordings>

12 www.bl.uk/learning/langlit/sounds/index.html

Lexical set**LOT****FOOT****FLEECE****BATH****Example in *Mr Tickle***

on, long, what, washed, somebody, stop, shop, doctor, dropped wood, took, looked, looking, could, stood, couldn't, butcher, pushes asleep, he, dream, reached, teacher, see, policeman, greengrocer past, fast, laugh, after, path, classroom, chance

Figure 1. Four of the lexical sets used in the *Mr Tickle* text

Objectives and methods

Map Your Voice ran from December 2010 to April 2011, while the British Library hosted a five-month exhibition, *Evolving English*, which examined the development of written and spoken forms of English over the centuries. Complementary to the online Map Your Voice project, the exhibition contained recording booths where visitors could record their accents. Over 14,000 examples were recorded in the booths by visitors to the Library's exhibition in London. The aims of the Map Your Voice project were to expand an existing set of recordings already in the British Library's collections by capturing new samples of *Mr Tickle* and other texts including from speakers beyond the UK and of all ages besides those of school ages.

Our methods used the online crowdsourcing techniques successfully deployed in a parallel British Library sound project, the *UK Soundmap* (Ranft 2011), that ran from June 2010 to June 2011 and aggregated from 390 different contributors over 2,200 samples of environmental sounds around the UK. We used Facebook, Twitter, blogs and other social media to generate and sustain interest in the *Soundmap*, winner of a UK social media award for innovation. Contributors primarily used mobile phones as recording devices using the free Audioboo¹³ application for iPhone and Android phones, or via Audioboo's website on a standard computer. The mobile Audioboo app automatically captures the recording location, time and date and allows users to add tags, titles and photographs. Combined with the audio file, we used these are to auto-generate a Google map mashup of recordings. That project took advantage of several trends: the ubiquity of mobile phones that have improved audio recording capabilities (there are more mobile phones in UK than the population); the fact that their owners carry their mobiles everywhere; and an ever-growing number of Audioboo users (who have created nearly 0.5 million sound recordings). We were able to tap in to this large active community of sound recordists, using the workflow for acquiring and archiving the Audioboo recordings described in Ranft (2011).

The *UK Soundmap* project successfully encouraged the capture of a new collection of a wide variety of everyday sounds, whether the sounds of nature, urban transport, street sounds or other public spaces, that together give an impression of Britain's sonic environment in the early 21st century, as chosen by contributing participants. For the *Map Your Voice* project, on the other hand, we needed to build on existing research collections of *Mr Tickle* recitations and other controlled vocabularies to improve the possibility of direct comparison between recordings, and by collating recordings worldwide. Each contributor was asked to tag their recording with the name of the place or region where they were brought up, their year of birth and their gender.

Interest in the project was triggered using conventional press and publicity methods coupled to the Library's complementary exhibition. It quickly generated considerable interest with many articles and media coverage, for example on BBC national radio stations¹⁴.

¹³ <http://audioboo.fm>

¹⁴ For example: www.bbc.co.uk/news/uk-11640951

Results

1,500 recordings were contributed to *Map Your Voice* in five months, via mobile phones and web browsers, primarily from North America and Western Europe with smaller numbers from Africa, Asia, Australasia and Latin America. We rejected 25% of the recordings offered. The reasons for rejection included the absence of location information, poor audio quality, or deviation from the prescribed text. The results were added automatically to a public website where contributors and anyone with an interest can play the sounds directly from an interactive map interface¹⁵ (Figure 2). Due to the large number of map points, we used Google fusion tables to create the map code rather than the standard Google map API that causes web page code bloat with many map points. The original audio files and metadata are being added to the British Library's collections and will be accessible permanently via a map interface on the Library's website at www.bl.uk/sound



Figure 2. The Map Your Voice interactive map

As an alternative to the *Mr. Tickle* text, contributors were also invited to recite a short sequence of six words: controversy, garage, neither, scone, schedule, attitude. These words were selected as they have known variations on a regional and global scale. Preliminary results from a random sample of 60 North American and 60 United Kingdom and Ireland contributors already have shown interesting patterns in the pronunciation of these words. For example the word *controversy* is pronounced CONtroVersy in the United States, whereas two thirds of UK and Ireland speakers say conTROVersy, stressing the second rather than the first syllable. Another example is the word *neither*, pronounced in one of two ways, with 75% of US speakers preferring *nee*-ther, 25% saying *nye*-ther. Across the Atlantic, UK and Ireland speakers in the sample showed a 50:50 split in the choice of pronunciation (Robinson 2011) (Figure 3).

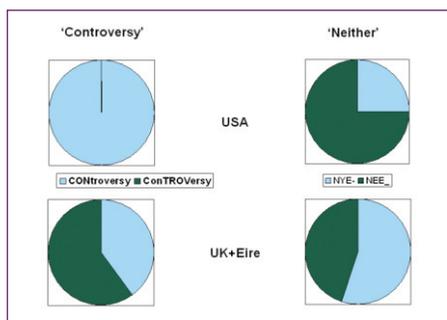


Figure 3. Pronunciations of controversy and neither

15 www.bl.uk/evolvingenglish/maplisten.html

Challenges

Several issues were addressed during project planning and assessing the results:

1. **Moderation:** the necessity for moderating contributions before publication to eliminate content with deliberate or inadvertent capture of copyrighted audio, defamatory words or obscenities. Moderating the sounds and giving prompt feedback to each contributor was the most time consuming part of the project.
2. **Sampling Bias:** compared to the general population, contributors tend to have a clear geographic and socioeconomic bias (middle class, metropolitan, British-born, familiar with IT and ownership of expensive smartphones). This caused a skewed representation of accents.
3. **Audio Quality:** the mobile phone recording quality is very variable, depending on the mobile phone type and the way it is used by contributors who may be relatively inexperienced as sound recordists.
4. **Metadata Quality:** the amount of user-generated accompanying information was the minimum to enable scholarly use of the aggregated recordings. More data would enhance the quality of the data but reduce the number of contributions.
5. **Map Interface:** The map positions shown on the public map at the time of the project showed the locations of the recordings, not necessarily the same as the locations of the dialects spoken. This caused confusion with many contributors and listeners.
6. **Authenticity:** a further consideration was that some users, rather than reading in their usual voice, would 'rise to the occasion' as if performing on stage or radio broadcast.

Conclusions

The process of making and uploading recordings for the collection website was largely managed by contributors, however manual filtering by the Library staff of the recordings was still essential for quality control. Library staff completed the process of archiving the audio files and importing the user-generated metadata into our catalogue. The metadata and audio quality was good enough to give a broad overview of regional variations and provide valuable examples for illustrating these. Their potential for more rigorous analysis will be limited by amount of metadata and variability in audio quality. Future projects could use a specially developed upload interface to ensure more controlled metadata and a higher audio quality level.

Nevertheless, Map Your Voice quickly showed its worth. It was simple to implement and cost effective. The research and educational value of the newly acquired recordings is proven; for example, samples of recordings are already being used by Sheffield University staff for their linguistics courses. Both the subject matter and the project methodology generated considerable interest and help to raise the profile of the sound collections at the British Library: the website alone received over 150,000 unique visits within five months. We could not have gathered such a data set any other way, and we hope to raise funding for a continuation of the survey in future. Other archives are encouraged to use this new technique in their own countries for the online acquisition of recordings submitted by volunteers.

Acknowledgements

Thanks to Jonnie Robinson, Lead Curator of Sociolinguistics at the British Library and to Chorion Limited for permission to reproduce the Mr Tickle texts in the exhibition.

References

- Hargreaves, R. 1971. *Mr Tickle*. London: Fabbri.
- Ranft, R. 2011. The UK SoundMap: an audio crowdsourcing experiment. *IASA journal*, 36:47-51



Robinson, J. 2010 *Why Mr. Tickle?*

<http://britishlibrary.typepad.co.uk/evolvingenglish/2010/11/why-mr-tickle.html>

Robinson, J. 2011. Eyether, eether; neyether neether — let's not call the whole thing off.

<http://britishlibrary.typepad.co.uk/evolvingenglish/2011/03/eyether-eether-neyether-neether-lets-not-call-the-whole-thing-off.html>

Wells, J.C. 2008. *Longman Pronunciation Dictionary*. (Third edition). Harlow, England: Pearson [check]

Wells, J.C. 1982. *Accents of English I: An Introduction*. Cambridge: Cambridge University Press

TAPE DEHYDRATION AS PART OF THE “JOURNALE” PROJECT: ON DEALING WITH STICKY-SHED SYNDROME

Stefan Kaltseis and Anton Hubauer, Österreichische Mediathek

The “Ö1-Journale” Project

The journal-radio-news program of the ORF radio-channel Ö1 (Austria I) is, without doubt, among the important acoustic documents of Austria’s recent history. After the broadcasting reform in 1967, the journals brought about a new form of radio news program. It was the result of a major change in information policy and radio legislation in Austria. Public discontent with the radio-news-program, which was very much influenced by the government, peaked in a national referendum in 1964, which requested that information in Austrian radio should be independent and a new legislation should confirm this. The radio news line “Journale” was a completely new type of broadcasting when implemented in 1967. Henceforth the central part of the radio-news for a day was, and still is the “Mittagsjournal”, the noon-journal, a one-hour daily news broadcast, except on Sundays, with detailed information on current events from national and international politics, economics and culture.

The “Journale” project focused on the cataloguing, digitization, long-term-preservation and online presentation of the “Mittagsjournal”. This centerpiece offers the largest spectrum of information. “Journale” are continuously moderated and the anchorman/woman offers a short introduction to each subject dealt with and to the contribution that normally follows. The contribution period has fluctuated over time, but nearly each one includes original audio of one or more people. Additionally moderators often interview people live in the studio. The “Journale” thus offers, day after day, a lot of “original voices”, and very lively and complex insights into the current situation and common attitudes. The flexible structure of the program made it easy to stress very up to date topics and to deal with most current affairs. Among the other formats from the “Journale” series, the morning and evening journals came into the project selection, though not regularly, to expand and deepen certain thematic topics of the project.

Overall, the project covers the period from October 1967 to May 1973 and from May 1976 to December 1989. The period from 1967 to 1973 consists entirely of footage from the ORF radio archive. The second period from May 1976 to December 1989 was covered exclusively by archival holdings of the Austrian Mediathek.¹⁶ The time-gap between the years 1973 and 1976 is mainly due to the strict financial and time framework conditions of the project.

The online platform

Over 4,000 tapes from the archive of the Austrian Mediathek have been digitized for the project. The archive format for long-term preservation is a high-resolution broadcast wave file; the format of the online platform is an mp3. Over 5,000 hours of air-time is accessible, free and unrestricted at www.journale.at, however, due to legal reasons, it is not available for download. Those 5,000 hours represent over 60,000 individual contributions to both Austrian and international issues on politics, economics and culture from 1967 to 1989.¹⁷

The initial situation

The tapes we had to deal with were AGFA PEM 369 semi-professional magnetic long-play tapes (polyester) on metal reels with audio content recorded at a speed of 3.75 inches per second. Out of the more than 4,000 digitized tapes, over 1,200 tapes were, and still are, affected by the hydrolysis effect, also known as the sticky-shed syndrome. All recordings from the year 1986 on are affected. This problem was more than a slight setback; the

¹⁶ <http://www.mediathek.at>

¹⁷ <http://www.journale.at>



whole project was in danger. Because of the rather tight budget and time frame of our project, a complicated and costly solution was not possible. The solution had to be technically as simple as possible, it had to have a serial processing character and it had to be cheap, really cheap.

Let's go a little bit deeper into the actual subject. Sticky-shed syndrome affected and endangered the whole workflow of the project, but what's it all about and how did it become recognizable in the first place?

Digitization at the Austrian Mediathek

After getting a magnetic tape out of the air-conditioned Archive, the first step is to evaluate the analog material, which means defining its tape speed, length and overall condition. The next step will be to write all this information in a special Job Database, which contains all the Metadata produced alongside the digitization process. Then the actual digitizing process takes place. In the case of the Austrian Mediathek that will be a 96 KH 24 Bit broadcast wave file with an mp3 at 128KB produced from this wave file. The magnetic tapes themselves are played on either Studer A807 or Studer A820 tape machines. Due to the fact that the magnetic tapes we had to deal with for the Journale-project were stored "tail-out", a method to prevent a print-through effect when playing them back which occurs after storing tapes for a long time, it was necessary to fast rewind them before we could start with the evaluation and the ensuing digitization.

The problem occurred

With tapes that were recorded up to 1986, the above procedure always worked. Suddenly, with tapes that were recorded from 1986, they exhibited the following effect: when trying to fast rewind the tapes we noticed that there was quite an abrasion on the tape machine. Besides that it seemed as if the tape was running clumsily over the tape head until it got stuck completely. It was not possible to work with audio material in this bad condition. So before we can talk about solutions to this it is necessary to ask why this effect occurs. What makes analog magnetic tapes become sticky?

Sticky-shed syndrome

Magnetic tape normally consists of a plastic backing — although the backing of older tapes was made out of paper and some of steel — which is covered by a layer of magnetic particles. This magnetic layer is made out of metal oxide that serves as the carrier of the information. In our particular case the sound is of radio news broadcasts. These two different materials, the plastic backing and the magnetic particles, are held together by a special binder that serves as a sort of glue. In the mid-1970s some of the major magnetic tape manufacturers started to change the mixture of this binder and, what they didn't know back then and what became only noticeable years later, is that the new formulation of this glue attracts moisture that leads to stickiness when trying to play back these tapes. It may also lead to the magnetic tape losing its contact with its plastic backing. In this context it's important to mention that not all tapes from a certain manufacturer are affected by Sticky-shed syndrome and among the ones that are in question there is very often a slight degree of difference in the severity of stickiness. What we can say from experience is that it is more likely that this has something to do with certain batches of production rather than with storing conditions or with a loose tape pack. It also has to be said that Agfa immediately changed the formulation of the binder as soon as the company found out, which to our knowledge was sometime in late 1990. The binder that was in use before contained polyurethane that soaks up water and causes the urethane to rise to the tape's surface. This attraction of moisture is also known hydrolysis.

Adverse impacts of sticky-shed syndrome

Normally it's not really obvious and, in many cases, one cannot really see whether a tape is sticky or not by just looking at it. You may have to put it on the tape machine and start playing it. This leads to yet another problem: not only is the tape itself decomposing in its parts and the quality of its recorded content getting worse, but this will also lead to the contamination of the tape machine itself. The binder and the magnetic coating of the tape pass over the tape machine and adhere to the tape head and the stationary guides as well as to other rotating parts of the machine that the tape touches.



The problem manifests itself mainly in two different ways: either the sound files become more and more dull and hollow, or the tape cannot be played back at all. Examples of affected sound extracts can be found on our website.¹⁸ Another typical symptom is a squealing noise when trying to play such a tape, sometimes oxide is even torn off the tape. Tape produced in the 1970s and 1980s is especially susceptible to break down as the oxide coating separates from the tape. The worst-case scenario is a completely contaminated tape machine plus damaged or torn magnetic tape. In addition to that it is obvious that a loss of magnetic particles means an irreversible loss of quality and information.

A look in the WWW

We were quite shocked about these effects when we first noticed them and they really endangered the whole online project. Nobody really knew how to deal with it so we started researching on the Internet only to find out that there were many archivists who were dealing with the same kind of problem. Other, and sometimes quite big, institutions were facing the problem of sticky tapes. Titles of articles we found on the web dealing with the problem range from "Tapes with Sticky-Shed Syndrome", "Dealing With Sticky-Shed Syndrome" to "Baking Audio Tapes Which Have 'Sticky-Shed Syndrome'" and the unusual yet accurately titled: "If I Knew You Were Coming I'd Have Baked A Tape!"¹⁹ The repeatedly given answer in online discussion forums about tape machines and magnetic tapes leads in one special direction: baking!

This is still a wide subject and there are many possibilities when it comes to "baking" analog magnetic audiotapes. One archivist even told us about a baking contraption made out of a

18 http://www.mediathek.at/oe1_journale/Wissenschaftlicher_Aufsatz_Eins/Tape_Baking.htm

19 <http://www.tangible-technology.com/tape/bakingI.html>

cardboard box, a hair dryer and a sugar thermometer. What is evident in the discussions on sticky-shed syndrome is that some audio archives use convection ovens that carefully create a uniform temperature with internal fans that evenly circulate hot air. We soon realized that this equipment is not cheap! Even for a rather big institution like the Austrian Mediathek with a collection of well over 400,000 objects this is a large purchase and who knows how often and how long we will need and use these ovens. But then there was another rather unusual solution: food dehydrators!

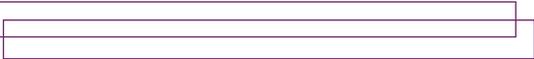
The practical implementation

For 150 Euro the Austrian Mediathek found drying devices made by a Swiss producer, including temperature adjustment and active warm-air circulation, a 12 hour timer and 3 drying trays. The 3 drying trays enabled us to “bake” 3 tapes in one dryer, with the same steady temperature and in an environment where the air is constantly circulating. So we started baking or rather drying our sticky tapes with these kitchen appliances.



6 dehydrators were in simultaneous use. 3 tapes per unit were dried at the same time, making a total of 18 tapes per day. We treated the tapes for 2 hours on one side, 2 hours on the other side, at 55 degrees Celsius. This process rendered playable tapes for digitization.

As for the digitization, with a workstation for 3 tape machines and a workstation for one tape machine provided by NOA audio solutions, the output in an 8-hour workday could have been up to 24 tapes, theoretically. However, dealing with problems such as broken playback equipment, updates and maintenance work of all kinds, the actual daily output was reduced to



somewhere between 12 to 20 tapes per day. This was not a major problem because the digital output matched rather well with the number of dried tapes. After a short period of time, between 10 to 14 days, the hydrolysis effect slowly starts to build up again. Within three weeks the effect is almost completely back. So it is not advisable to wait longer than 10 days after drying a tape to digitize it. What we dried one day, we digitized the following day.

The permanent failure of the years 1986 to 1989 would have introduced an additional time gap for the project. The whole project goal was in severe danger. Using these domestic dehydrators provided a simple, time-effective and cheap solution. In total, the dehydrators cost less than €1500, a financial burden, which was borne by the Austrian Mediathek itself, because the equipment will provide continued good service when needed.

Experimental studies

In preparation for this article we once again picked up ten of the tapes from the archive that we knew were sticky and started to digitize them right away without further cleansing or “baking”. The results were again, after about two years of not having touched them, disillusioning. All of the tapes sounded hollow, some of them were not playable at all and some made the tape machine stop almost immediately. Another point to mention is that the azimuth value which should be around zero is much harder to control with these tapes than under “normal” conditions.

Next, we dehydrated the tapes in the food dehydrators again for 2 hours on each side at a temperature of 55 degrees Celsius (131 Degrees Fahrenheit) and digitized them again. All of the tapes were playable without contaminating the tape head or other parts of the machine; none of them showed signs of clumsiness and even the azimuth value was easier to control than before. So these tapes would have been able to be digitized without any further abrasion of the iron oxide coating or even risking a tape-rupture. And they sounded ok!

But we waited for another three weeks before we digitized the tapes again to see if hydrolysis was already starting to affect the quality of the tapes. And indeed, it did. The tapes again showed signs of stickiness and a loss in quality when played back. Some of them were not even playable at all! After dehydrating them a second time, they were without any problems ready to be played back and transferred to a digital device.

Lessons learned

What insights concerning the handling of tapes affected with sticky-shed syndrome did we gain during the implementation of the “Journale”-project? The main advantage of dehydrating sticky tapes as described above is apparent. First and foremost, it is cheap and effective. Furthermore, we think it is a sensitive way of getting rid of that unpleasant effect. Even for private collectors who want to digitize their audio material, this is an affordable and reasonable way of solving the problem. The quality does not audibly get worse and, in our experimental case, this can be said even after dehydrating the tapes more than once (although we do not recommend that). Even the frequency level shows no differences compared with an unaffected reference tape.

Some advice

Before we come to some final remarks we would like to give you some advice in case you want to start dehydrating your analog tapes in anticipation of digitization.

First of all, do not dehydrate magnetic tapes on plastic reels but rather transfer them to metal hubs first. Although the chances are slim and some experts say that this will not be the case at these temperatures, plastic reels can warp and become deformed by heat.

Do not rewind or play the tape before you bake it, unless of course you want to exchange the plastic reel for a metal one, this will just further add to more damage on both the tape and



the tape machine. The problem here is that you do not really see whether a tape is affected by sticky-shed syndrome before it is placed on the machine and let run for a few seconds. Don't go any further than this and if you are not sure whether it is an affected tape or not, rather bake it before you risk destroying it. However, another proposition that we read about quite frequently contradicts this advice, which is to try to achieve an even pack before baking. Baking uneven tapes may damage the edges because they get too warm and this causes lasting artifacts in the stereo sound. The decision is yours as to whether it is better to have an even and tight pack or no further abrasion of oxide from the tape when you decide not to fast wind it before the dehydration process.

Furthermore make sure you are dehydrating your tapes at a stable and steady temperature as mentioned above. It is necessary, therefore to have a thermostat, which is normally included in a standard food dehydrator. Afterwards, return the tape to its box and let it cool for approximately 24 hours as a precautionary measure. The fact is that not all playback machines are equally gentle and a tape that is still warm is more likely to stretch than one at normal temperature.

Our experience showed that tapes could be baked more than once. Nevertheless do not attempt this until unless absolutely necessary: try to digitize the baked tape within the 2 week time frame. The quality definitely does not getting better if you perform the baking procedure over and over again for pure listening purposes.

In the discussion following a presentation at the last IASA conference held in Frankfurt earlier this year a colleague recommended wrapping the tapes in plastic bags and including silica gel to absorb moisture. So far we haven't attempted this technique; however we consider this to be another possible way of prolonging the positive effects of the dehydration of sticky tapes.

Final remarks

In contribution to a research project of the German Rathgen Institute in Berlin, a scientific institution for conservation and restoration, we sent some tapes from the archive of the Austrian Mediathek for further analysis of the impact that sticky-shed syndrome has on the chemical level for magnetic tapes. Experts at this research institute are trying to find out more about what actually leads to hydrolysis in magnetic tapes and what happens during the dehydration process in order to reverse sticky-shed syndrome. With this information it might be possible to figure out an ideal temperature and duration of dehydrating magnetic tapes affected by sticky-shed syndrome. Results of this study are expected soon and will be published on the Rathgen Institute website.²⁰ We are excitedly looking forward to hearing about their findings in order to gain an even better understanding of what actually happens during the process of dehydration.

There are still a lot of unanswered questions about how and why some tapes tend to become more sticky than others and why this effect does not hit some at all or only to a smaller degree. Let's hope that research studies as the abovementioned may cast a brighter light on that problem. Meanwhile, what we tried to show in this paper was how a renowned and representative audiovisual archive like the Austrian Mediathek is dealing with the problem of magnetic audio tapes affected with sticky-shed syndrome in a down to earth but nevertheless effective and — as far as we can tell so far — harmless way.

Sticky-shed syndrome is not a question of the correct way of archiving magnetic audiotapes. Even in the best and most ideal preservation environments, tapes can get sticky and binders can start to decompose. The problem we are dealing with here is inherent in the material and its chemical composites itself. Our experience shows that with the described method so far no audible aggravation of the audiotapes is recognizable and it is possible to preserve and transfer analog audio content into the digital era, where, yet again, new forms of problems are waiting to be solved.

20 <http://www.ilkarde>

STORAGE STRATEGY TOOLS

Matthew Addis, Mariusz Jacyno, Martin Hall-May, IT Innovation Centre, University of Southampton, UK, and Richard Wright, Archive Research, BBC Research and Development, UK

Introduction

This article describes the challenges of selecting a storage strategy for long-term retention and access of digital content and how some of the tools developed in the PrestoPRIME²¹ project are designed to help.

Long term storage of digital content can be a significant cost, although not typically the biggest cost factor for preservation and access — digitisation, cataloguing and rights management often dominate for many audiovisual archives. In many cases, storage is also a cost that falls rapidly, with some claiming that storage is essentially becoming ‘free’, at least in comparison with other costs. However, currently, storage is expensive enough to be a significant part of preservation budgets and as such requires due attention. Storage is also a recurring cost. Worse still, in the case of digital content stored on IT systems, storage is a recurring cost where payment has to be kept up or otherwise content is rapidly at risk of being irretrievably lost due to obsolescence and storage failures. This is where the real ‘cost’ lies: the cost of doing nothing or the cost of doing it wrongly — both result in loss of content, i.e. the ‘cost of loss’.

Storage is not about giving digital content a safe harbour where it can live out its days. Given the ever increasing drivers for making archive content rapidly and easily accessible, and with a ‘holy grail’ of having ‘everything online’ with ‘instant access’, storage should be considered as a fundamental component of enabling access as well as safe keeping. For many audiovisual archives the reason content is kept is so it can be reused, repurposed and enjoyed again. The type of storage used to store content has a big impact on the accessibility of content being stored: compare film in a freezer with disk servers on a high-speed network. Yet the types of storage that provide the fastest access are not necessarily the ones that have the lowest cost of ownership, nor the highest degree of safety for the content within them. You can strive for low-cost, high-safety, and fast-access — but it’s not possible to achieve all three at once!

Here lies the dilemma. There is a trade-off between cost, ease of access, and safety of digital content when using any type of storage. There is no ‘one size fits all’ solution. Neither is there a ‘one size fits all’ set of needs from archives. The challenge is one of combining different types of storage and matching the combined characteristics to an individual archive’s needs. This means matching storage strategies to the needs of an archive both today and in the long-term. We attempt to tackle this challenge with our simulation and modelling tools, in particular by allowing the costs and risks to be compared — which we call the ‘cost of risk of loss’ approach.

State of the nation

There is an increasing body of work on the total lifetime costs of storage [1][2][3]. Costs include power, space, cooling, maintenance, migration and management [4] in addition to the more obvious costs of media and the physical servers or shelves used to store it. Some of the cost elements fall rapidly over time, e.g. following Kryder’s law,²² although there is concern on how much longer this will hold.²³ Some costs, for example power, may actually increase, although the net effect is still downwards when increased storage densities and efficiencies are taken into account. If past performance is an indicator of the future, then the Total Cost of Ownership (TCO) of storage will continue to fall for at least the next decade, and at a significant rate. For example, the Internet Archive can be used to look back at the costs of online storage service providers (e.g. Amazon S3) over time, with the charge per TB per year seen to

21 PrestoPRIME <http://www.prestoprime.org/>

22 http://en.wikipedia.org/wiki/Kryder%27s_Law#Kryder%27s_Law

23 <http://www.storagenewsletter.com/news/disk/hdd-technology-trends-ibm>



halve every 2-3 years [1]. This is slower than the rate of fall of underlying media costs, e.g. the cost of a hard drive in a local electronics store halves every 18 months or less. The slower rate of fall represents the effects of all the other cost factors that make up TCO (physical media can be as little as 10%)[4]. So confident are some organisations [5] in this trend of falling costs that that they will accept a one-off payment today in exchange for a commitment to store content indefinitely — a so called ‘forever price’.

There is also a significant body of work on the reliability and safety of storage for digital content [6,7], much of it focusing on IT technology, e.g. Hard Disk Drives, data tape and optical discs. This work is well known to the vendors of such storage, as evidenced by increasingly complex systems to guard against failures, e.g. double or triple parity RAID arrays [8] and tape libraries with automated data integrity checking.²⁴ There is some debate over the value of the metrics used by the storage industry for safety, e.g. as highlighted by the paper “Mean time to meaninglessness” [9] with better ways to measure safety also proposed - but this only serves to highlight that IT storage systems aren’t fundamentally reliable ways to store data in the long-term. That is unless a suitable active data integrity management strategy and set of processes are put in place. This active management approach can now be seen at all levels including archive asset management systems,²⁵ file systems²⁶ as well as inside storage systems as already discussed.

Storage reliability and safety is faced by the people using these storage systems ‘day to day’, e.g. IT departments and their systems managers — although this is often learned the ‘hard way’ through experience of loss rather than knowing what best practice to apply upfront. The main problem is that this knowledge of storage reliability is relatively unknown to the archive community — generally speaking, good practice has yet to be widely established when using IT systems for digital preservation of AV content. Worse still, techniques commonly used by archivists for storage in the ‘analogue domain’, e.g. ‘putting media items on shelves’, are often assumed to translate into the ‘digital domain’. In fact the opposite is true. Storage of digital data requires active management and a plan of regular interventions (refreshes, migrations, integrity checks and repairs) to ensure data remains intact and the systems are in good shape.

Whilst information exists separately on both the costs and risks of different types of storage, a lot less work has been done to combine the two, especially in the context of storage being an enabler for access to content. Existing work also tends to focus on single storage systems rather than their combination when adopting good preservation practice of ‘multiple copies in multiple locations, using diverse technologies’. We attempt to address this gap in our work.

Cost, access and safety don’t go hand-in-hand

There are many different approaches to storing digital data. Each has its own strengths and weaknesses in terms of safety, ease of access, lifetime and cost. For example, long-lived storage technology includes printing bits to film²⁷ but costs per TB are high and access requires a film scanner. IT storage technology includes data tape or HDD, with data tape offering lower TCO at high data volumes and a good level of safety, but not the rapid ‘random access’ that HDD servers can afford — but for HDD servers this increased ease of access comes at the expense of increased cost and shorter time to obsolescence. Making more copies of content is of course one strategy for increasing safety, e.g. an extra copy in an offsite deep archive, which is essential to guard against some catastrophic types of failure, but does so with significant extra cost. The content itself can be encoded to make it more resilient to failures in storage [10], but this adds complexity and potentially a different set of risks from using a format that is not ‘standard’ and hence itself at risk of becoming obsolete. Error correction is commonplace in almost all types of digital storage and works tirelessly behind the scenes to correct a multitude

24 Examples in 2011 include SpectraLogic and Quantum adding this feature to their libraries/software stack

25 For example, DIVArchive from Front Porch Digital

26 For example, ZFS. <http://en.wikipedia.org/wiki/ZFS>

27 For example, DOTS from Cinevation. <http://cinevation.net/>

of sins. Further layers of error detection and correction can be added to storage to catch and repair problems (e.g. the use of checksums and ‘scrubbing’), but this adds additional cost. Errors can also be concealed. When it comes to errors, it is not actually lost bits and bytes that typically matter — rather the impact they have on the ability to use the content. Here it is possible to conceal this impact, e.g. artefacts introduced by block-level corruption in a video stream can be interpolated from adjacent frames and often made visually acceptable.

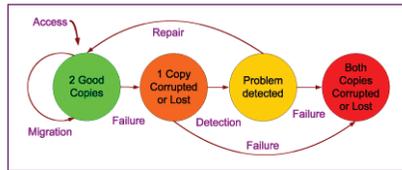


Figure 1

Figure 1 above presents a conceptual model for comparing and combining different storage approaches. With reference to the diagram, the bedrock of data safety is to keep multiple copies of content (green circle), typically by using different storage technologies and in different locations, and ideally operated by different people. This guards against major risks (e.g. it allows disaster recovery), but it also guards against unanticipated problems with individual technologies and processes (i.e. it ensures that the eggs are not “all in one basket” at any level). For each copy, there is the need to regularly migrate each component of the technology stack (hardware, operating system, management software, file formats, etc.). One or more of the copies is also used to serve access requests. However, there is always the chance that one of the copies could be damaged or lost due to some form of failure in the system (orange circle). Only after this problem is detected (yellow circle) can any action be taken to repair, replace or conceal defects in the damaged or lost copy. If at any time before this process is complete something happens to the second — and only remaining good copy — then there is a risk that both copies could be permanently lost or damaged (red), and the content could be lost. Depending on the type or damage or loss, there is the option of doing partial repairs, concealing defects (a common approach in digital video), or reconstructing a good copy from parts taken from two damaged copies. All these options have additional costs and benefits. The rate at which transitions happen between the states dictates the length of time during which content is at risk. Every transition also has a cost. Therefore, whilst superficially quite simplistic, the model above provides a framework for the total cost and total risk to be assessed and storage strategies to be compared. A detailed qualitative comparison of a wide range of storage techniques is available in an article by the authors in the SMPTE digital imaging journal [10], also available online in extended form as a public PrestoPRIME report [11].

Tools

Our approach to quantitative comparison of the costs and risks when using storage starts with a simple but flexible storage model (Figure 2) that consists of a set of storage systems (e.g. tape library, HDD server). Each storage system has the function of accepting files for storage (writes), returning files from storage (reads) and storing the data inside the files using some form of physical media (hard drive, data tape, optical disc etc.). The model includes a ‘controller’ that mediates access to the underlying storage. The controller could be a person, i.e. manual operation of ‘media on shelves’ archive, or the controller could be automatic, e.g. storage management software operating a tape library.

The reason for choosing a simple model is so it can be applied to both automated hardware/software, e.g. a HDD server, and to more manual processes, e.g. data tapes on shelves where archive personnel put new tapes onto shelves and retrieve existing tapes to serve user access requests. When writing or reading files, various operations may be applied, e.g. encoding or applying error correction. Depending on the system being modelled, this could be by firmware on a HDD, the RAID controller in a HDD array, integrity management in a ZFS filesystem, manual integrity verification by an operator, or a combination of all of these. Likewise, various failures or errors could occur, both latent (failures that happen silently [12]) or extant (failures that are immediately detected). Failures can range from 'bit rot' in a HDD system through to accidental damage from manual handling of discrete media, e.g. data tapes. These failures can happen (a) when data is written, (b) when data is read, and (c) when the data on the physical media is in effect 'doing nothing'. In our model, these are all represented through error rates for read/write/store actions. The actions each have a cost, which forms one part of the associated cost model (e.g. one-off ingest cost per file when adding it to a storage system, access cost per file incurred each and every time it is retrieved from the storage system, and the on-going storage cost per file when it is inside the storage system).

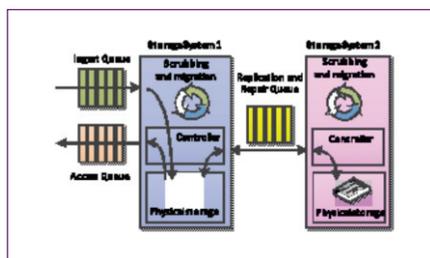


Figure 2 Archive storage strategy

In the model, one or more storage systems are then combined into a storage strategy. Figure 2 shows just one example that combines two storage systems. The storage strategy includes ingest for new file arrivals and access queues to serve user requests for content. The strategy determines how files are allocated to storage, how they are replicated, and how they are repaired if there are failures. Resources can be allocated to serving ingest, access and copy operations as well as for activities within each storage system, e.g. integrity checking and repair. It is often these resources that are limited, especially at peaks of workload, so limits can be set in the model to allow investigation of what happens when resources run out.

We have made two tools available²⁸ that allow projections of cost and loss over time. The first tool takes a very simple approach that allows storage systems to be characterised in just 4 parameters (cost of storage, cost of access, latent corruption rate, access corruption rate). Two storage systems are combined into a storage strategy (2 copy model). This includes parameters for the frequency of checking and repair of the data (scrubbing), the number of files to start with and how fast this increases, how often access to these files takes place, when migration is needed and how costs change over time. The tool comes pre-loaded with parameters based on real-world storage systems, with some examples given in Table 1 for a scenario of storing 25GB files (e.g. approx. 1 hour of SD video at 50Mbit/sec). Full derivation of these values can be found in [11] and they of course vary over time and with the size of the files.

28 <http://prestoprime.it-innovation.soton.ac.uk/>

	Latent corruption rate (files)	Access corruption rate (files)	Storage costs (Euro per GB per year)	Access costs (Euro per GB)
HDD online (server)	1 in 750	1 in 500	1	0.1
HDD offline (shelves)	1 in 100	1 in 500	0.1	5
Data tape online (library)	1 in 100,000	1 in 10,000	0.5	0.1
Data tape offline (shelves)	1 in 10,000	1 in 1,000	0.05	5

Table 1 example parameters for storage of 25GB files

The simulation then projects the number of files that are ‘alive’, ‘at risk’ or ‘lost’ over time (lost means that both copies have become corrupted, at risk means that one copy has become corrupted, alive means that there is at least one copy that hasn’t been corrupted). An example is shown in Figure 3 for an archive that stores its contents using HDD on shelves. For simplicity, the number of files in the archive remains constant and there is no regular ‘scrubbing’ to check integrity (due to the costs of doing this for a HDD on shelves model). The number of files at risk climbs year on year, as does the annual number of files lost, until a migration point which provides the opportunity to detect and repair losses.

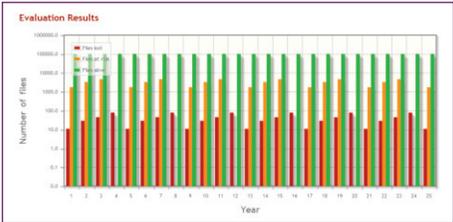


Figure 3 Files at risk and lost for a 25 year projection for a HDD on shelves strategy

The corresponding costs are shown in Figure 4. As with the risk/loss projection, cost per year is fixed to make the graphs easier to understand for the purposes of this paper. The cost is broken down into each storage system (with one higher than the other due to user access to content being delivered through storage system A and not system B). The spikes in cost every 4 years correspond to migrations. It is then possible to use the tool to adjust parameters and look at the consequent impact on costs and loss over time. In this way, different storage strategies, e.g. tapes and HDD in servers/libraries can be compared with a ‘media on shelves’ approach.

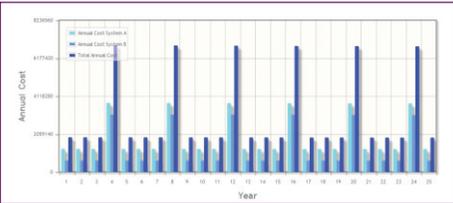


Figure 4 Cost for a 25 year projection

It should be clear that the projections from the simple model are exactly that — simplistic. They do not provide the detail or accuracy for use in business planning or day-to-day operation. The objective is however only to provide a simple comparison of some of the strategies that could be used for storage. In this respect, the tool is a useful educational aid in illustrating the importance of an active storage management approach if data integrity is desired over the long-term.

To allow a more in-depth analysis of storage strategies, the second tool developed provides an interactive simulation tool based on a discrete event simulation approach. During the simulation, time ticks away (e.g. 1 second of the simulation might correspond to 1 week in the real world) and events are generated (e.g. corruption of files in a storage system, requests to access a file, new files to be added to the archive). These events then trigger actions, e.g. a copy/repair process, which is then added to the queues of the storage systems involved. A storage system processes items in its queues according to how much resource it has available (e.g. serving access requests sequentially or in parallel). The available capacity of the resources used by each service determines how many items are processed for each tick of the clock, and at what cost.

The user can interact with the simulation as it progresses, e.g. changing the amount of resources available or changing the policy for data safety (e.g. making more copies or checking them more often). In this way, the user is in effect playing a game that helps them understand how to react to and manage events that they might see in practice when operating a real system. For example, there is also an option to simulate 'disaster scenarios': rare but catastrophic events where large fractions of the storage become temporarily or permanently unavailable.

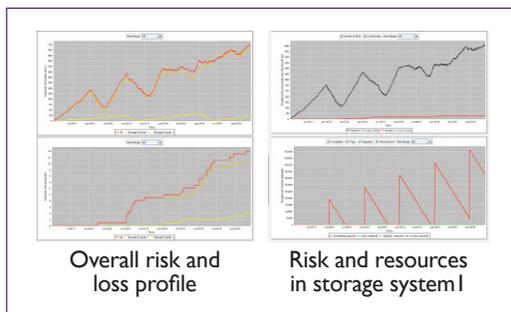


Figure 5 Interactive simulation of a combined HDD and data tape storage strategy

An example of a simulation is shown in Figure 5, again for a 2 copy model, but this time modelling a combination of HDDs in a server and tapes in a library. Periodic scrubbing is included as well as migration, and there is growth in archive content over time with a proportionate amount of access requests to go with it. All starts well with a low rate of content loss and regular integrity checking well within system capacity. However, as the archive grows there is an increasing access burden and more content to manage. Priority is given to ingest/access over background scrubbing with the result that integrity management starts to suffer and eventually runs out of resources with a consequent increase in loss of content. At any point during the simulation it is possible for the user to adjust the resources in the system in order to counteract this effect, and hence estimate how the capacity of the system will need to be extended over time. The example above is intended to be illustrative and the tool allows much more sophisticated simulations to be run, recorded and rerun.

Repeated running of the simulation with variation in input parameters allows a 'map' to be created of the cost/risk 'landscape' for a given storage strategy.

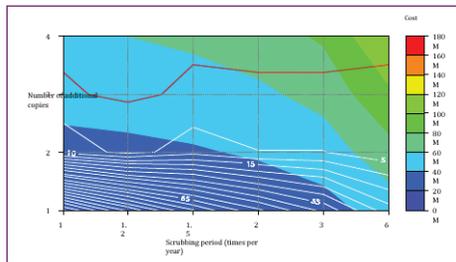


Figure 6 Cost of risk map

An example map is shown in Figure 6. This shows a single storage system where the number of additional copies of a file stored in that system and frequency of integrity checking (scrubbing) both impact on both the cost and the risk of file loss. The boundary between adjacent coloured bands represents configurations of equal cost. The white contour lines are lines of equal risk of loss and represent the peak number of assets where only one uncorrupted copy is left (which is typically a transient position because scrubbing picks up and repairs the corruption). The red contour line is the boundary between the probability being in favour of at least one file loss or no file loss over a 10 year period. Increasing the number of copies reduces the risk of loss, as expected, but also increases the cost because more storage capacity is needed. Increasing the scrubbing frequency also reduces the risk of loss, but again increases cost because of increased access to data and equipment needed to compute checksums for a larger volume of data. For the particular parameters used for the storage system shown, it is more cost effective to add more copies than it is to scrub them more often and there is a ‘sweet spot’ that balances the two to give the lowest cost of zero probable loss over 10 years. Different storage strategies have different balances, which is why a tool to allow the trade-offs to be analysed on a case-by-case basis is so important. If zero loss is required then it also allows the area on the landscape to be found that is sufficiently ‘far’ from risk contours to give some margin for error between the simulation, e.g. because of lack of precise input data, and what happens in the real world, e.g. because things rarely go according to plan!

Next steps

The tools described in this paper are still in their early days. There is much that could be done to add functionality and improve them. The next steps of our work are three-fold. Firstly, we are investigating how we can best validate the models, i.e. give confidence in their results. This can be achieved in several ways, e.g. (a) comparison with analytical approaches for simple test cases that are also tractable this way, (b) comparison with the findings of field studies and real archive experience — although this is hard because whilst some data exists it is insufficient to validate all aspects of the model, and (c) validation of the implementation by allowing independent inspection of the source code and design of the model to pick-up potential bugs — which is one reason why the tool is available as open-source and we are transparent with our design assumptions and testing. Secondly, we are planning to develop support in the model for higher-level preservation functions, e.g. transcoding of file formats (another type of migration) and repair/concealment of the impact of corruption at the AV level, e.g. mapping the effect of data corruption onto different video formats and how the corruption could be concealed or repaired. Thirdly we are looking at ways of making the tools simpler to use. The simpler the tool, the less time and investment it takes to learn how to use it and get the required results. However, if this is achieved by making the tool too restricted then the model and the real world diverge and the results of modelling are less valuable. There is a balance to strike between model complexity, model accuracy, ease of use and what users of the tool are willing to invest in terms of time and effort learning how to use it.

Acknowledgements

PrestoPRIME is an EC supported 7th Framework Programme ICT project (FP7-231161) coordinated by INA (Institut National de l'Audiovisuel) in France. Partners include BBC, RAI, ORF, B&G and others. For further information see www.prestoprime.org

References

- [1] Matthew Addis and Richard Wright (2010). PrestoPRIME Deliverable D2.1.1 “Audiovisual preservation strategies, data models and value-chains”.
<http://www.prestoprime.org/project/public.en.html>
- [2] David Rosenthal's Blog. See post “Modelling the Costs of Long Term Storage” posted on 27 September 2011. <http://blog.dshr.org/> Consulted 12 October 2011
- [3] Richard Wright, Matthew Addis, Rajitha Weerakkody (2010). “Century Store, Real Options, Real Costs”. Published at the AMIA/IASA 2010 Joint Conference, 2-6 November, Philadelphia, USA.
- [4] Barroso, L.A. and Holze, U. (2009). The Datacenter as a Computer: An introduction to the design of warehouse-scale machines. Google Inc. Synthesis Lectures on Computer Architecture no. 6. Published by Morgan and Claypool.
- [5] Serge J. Goldstein, Mark Ratliff (2010). DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data. Office of Information Technology, Princeton University. August 27, 2010.
- [6] Elerath, J. (2007). Hard Disk Drives: The Good, the Bad and the Ugly!, Queue 5, 6, p28-37.
<http://doi.acm.org/10.1145/1317394.1317403>
- [7] Jiang, W. et al. (2008) Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. FAST '08. <http://www.usenix.org/events/fast08/tech/jiang.html>
- [8] Adam Leventhal (2009). Triple Parity RAID and beyond. ACM Queue vol. 7, no. 11. <http://queue.acm.org/detail.cfm?id=1670144>
- [9] Kevin M. Greenan, James S. Plank, Jay J. Wylie (2010). Mean time to meaningless: MTTDL, Markov models, and storage system reliability. In proceedings of The 2nd Workshop on Hot Topics in Storage and File Systems (HotStorage '10), June 22, 2010, Boston MA, USA
- [10] Matthew Addis, Richard Wright, Rajitha Weerakkody (2011). “Digital Preservation Strategies: the cost of risk of loss for AV Content”. Jan/Feb 2011 edition of the Motion Imaging Journal of the Society of Motion Picture and Television Engineers (SMPTE).
- [11] Matthew Addis, Mariusz Jacyno (2010). PrestoPRIME deliverable D2.1.2 Tools for modeling and simulating migration-based preservation. <http://www.prestoprime.org/project/public.en.html>
- [12] David Rosenthal (2010). Keeping bits safe: how hard can it be? ACM Queue vol. 8, no. 10
<http://queue.acm.org/detail.cfm?id=1866298>

FONOTECA NACIONAL OF MÉXICO LAUNCHES THE PLATFORM FOR SOUND AND AUDIOVISUAL HERITAGE PRESERVATION OF IBEROAMÉRICA

Álvaro Hegewisch, *Fonoteca Nacional (IASA membership vice-president) and Xilonen Luna, Comisión Nacional para el Desarrollo de los Pueblos Indígenas*

The richness of sound and audiovisual collections in Latin America is evident, unfortunately so is their imminent disappearance. The definition of policies and joint actions is crucial to generate conditions of preservation in the region. Many efforts have been made, but usually these efforts approach only one part of the process and result in isolated and sporadic actions.

The creation of the Fonoteca Nacional in December 2008 gave México the leadership in the preservation field and in the promotion of sound heritage not only within the country but in the Latin American region as well.

Since 2010 our challenge has been the consolidation of this leadership to create the Program for the Preservation of Sound and Audiovisual Heritage of Iberoamérica. We are convinced that if we want to have success in the preservation of our sound and audiovisual archives, we need to join *experience, knowledge* and *resources* in a full process model, with representative institutions of the Iberoamerican region and, of course, with international organizations, mainly with IASA and FIAT.

Drawing on their experience and awareness of this situation, the Fonoteca Nacional de México, with the collaboration and support of IASA and FIAT, promoted a Latin American Meeting of Sound and Audiovisual Archives in August 2010. The purpose of this meeting was to form a working group, representing countries of the region and to define an agenda that would develop *integral projects to rescue collections at risk*.

Conceptual references, priorities and a working agenda were defined in this meeting. The agreements are:

1. Establishment of a basic diagnostic questionnaire
2. Development of a didactic guide
3. Development of a Glossary
4. Production of a video tutorial
5. Model process of a basic preservation plan for archives in risk
6. Pilot Project in México with Indigenist Radios.

As a result of this agenda, the Fonoteca Nacional developed the web-based Platform for the Preservation of Sound and Audiovisual Heritage in Iberoamérica, (www.fonotecanacional.gob.mx), to make available simple and effective tools for the safeguarding of archives from the region, in our language and according to particular circumstances.

Fonoteca Nacional **iasa** **CONACULTA** **MÉXICO**
es cultura

Plataforma Iberoamericana para la preservación del patrimonio sonoro y audiovisual

QUESTIONARIO PARA LA IDENTIFICACIÓN Y DIAGNÓSTICO PRELIMINAR EN RIESGO

GUIA DIDACTICA INTERNACIONAL DE ARCHIVOS SONOROS Y AUDIOVISUALES

BIBLIOGRAFIA Y DOCUMENTACION ESPECIALIZADA EN REFERENCIA A LOS ARCHIVOS SONOROS

FORMACION Y CAPACITACION PRESENCIAL, ONLINE Y DOCUMENTACION EN LINEA

Con la apertura de la Fonoteca Nacional del Consejo Nacional para la Cultura y las Artes, México emprendió acciones contundentes a favor de la preservación del patrimonio sonoro nacional. Por ello y gracias a la colaboración de la Asociación Internacional de Archivos Sonoros y Audiovisuales (IASA), la Fonoteca Nacional de México ha creado la **Plataforma para la Preservación del Patrimonio Sonoro y Audiovisual de Iberoamérica**, con el propósito de poner a disposición de los actores de la región herramientas para la salvaguarda de este patrimonio en riesgo.

Una de estas herramientas es el **Cuestionario de diagnóstico del patrimonio sonoro y audiovisual**, que la Fonoteca Nacional de México publica en esta Plataforma con el objetivo de identificar y conocer el estado de conservación en el que se encuentran los archivos sonoros y audiovisuales en riesgo en Iberoamérica.

Además, en esta plataforma la Fonoteca Nacional de México ofrece documentación, formación en línea, intercambio de experiencias y asesoría para generar un plan de acción inmediata, en especial para aquellos archivos que se encuentran en riesgo inminente.

De esta forma, la Fonoteca Nacional de México promueve la colaboración internacional a favor de la preservación del patrimonio sonoro y audiovisual de Iberoamérica.



One of these tools is the Diagnostic questionnaire of audio and audiovisual heritage. This tool's purpose (which looked at the experiences of TAPE in Europe and the ICRT questionnaire in Cuba) is to identify and understand the condition in which there are audio and audiovisual files at risk in Iberoamérica. The questionnaire has guides (instructional videos and a glossary of terms) which may be consulted on the website of the Fonoteca Nacional of CONACULTA (www.fonotecanacional.gob.mx). This diagnosis will result in an immediate care plan, especially for those files that are at risk of being lost.

It also seeks to promote the integration of an Inventory of Sound and Audiovisual Heritage in Iberoamérica with homologated and approved IASA standards, coordinated by México through the Fonoteca Nacional of CONACULTA.

In addition, this platform provides specialized documentation and literature, in Spanish, about the sound and audiovisual archives, online training, exchange of experiences and advice on creating a plan for immediate attention, especially for those files that are at risk of being lost.

The first goal is to know the conservation condition of the sound and audiovisual files in order to identify archives at risk in the region and implement plans for emergency care.

At the same time we are promoting a “*training for trainers*” program with the National Sound Archive in México, the Phonogrammarchiv of Austria and the Swiss National Sound Archive. In the medium term, this program seeks to:

1. Promote the integration of an inventory of the Iberoamerican Sound and Audiovisual Heritage, with standards approved and endorsed by IASA and FIAT, and coordinated by México, through the Fonoteca Nacional del CONACULTA.
2. Develop a “training for trainers” program, in Spanish, for technical education (basic and specialized) in sound and audiovisual preservation.
3. Establish the Fonoteca Nacional de México as a knowledge resource center within the region.
4. Promote the Iberoamerican IASA branch.

The Fonoteca Nacional provides advice and training in preservation areas in order that each archive has the necessary tools to generate an action plan for the immediate attention of its sound and audiovisual heritage at risk. Until now the Fonoteca Nacional has provided advice to Colombia, Costa Rica, Honduras and México.

Recovery of sound documents at Radio XEZON from the Sistema de Radiodifusoras Culturales Indigenista (Indigenous Cultural Radio Broadcasting System) and from the Comisión para el Desarrollo de los Pueblos Indígenas (National Commission for the Development of Indigenous Communities)

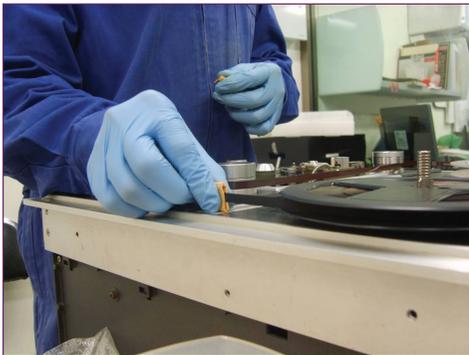
In October 2010, according to the last agreement point in the agenda, the Fonoteca Nacional de México and the Comisión para el Desarrollo de los Pueblos Indígenas started a joint pilot project in México for the safeguarding of one specific sound heritage at risk, namely, the sound archive of XEZON La Voz de la Sierra de Zongolica, Veracruz.

As a result of this work, 410 tapes from the Radio XEZON The Voice of the Sierra de Zongolica were received at the Fonoteca Nacional, in order to submit them to an action plan for rescue and conservation. These sound documents are of great significance to the *Nahua* people of the Sierra de Zongolica for their historical content relating to traditional music and oral tradition.

The process began with the diagnosis made with the chiefs from the archives of the Radiodifusoras Culturales e Indigenistas through the application of a questionnaire and the provision of a basic course to the Radio XEZON staff. Thanks to this process we were able to identify the materials that had a biological deterioration caused by fungi and crystallized adhesive in the tapes. Within the emergency care plan for its conservation, the

following measures were implemented:

- Replacement of boxes and reels that suffered from mold damage
- Stabilization of the media in the transit vault with temperature and humidity control
- Allocation of inventory numbers for both boxes and reels, in order to keep the original information of the documents
- Digitizing. This process took into account several factors such as: material and brand of tape, winding, state of the reel, original defects and defects caused by the state of preservation.



The rescue of the Sierra de Zongolica community oral history tapes is an example of “good practice” in implementing a model of identification, diagnosis and care of collections at risk with multiple benefits for the sound heritage of the Sistema de Radiodifusoras Culturales Indigenistas.



Conclusion

With the Iberoamerican Platform, Fonoteca Nacional de México promotes the identification, diagnosis and care of sound and audiovisual archives at risk in Iberoamérica. Now the challenge is to have more strategic alliances like CDI in México, IASA and FIAT, and the participation of more archives in the region to achieve the following main goals:

1. Integration of an inventory and a network of Iberoamerican archives in the region
2. Certification of the training of trainers program with direct benefits in the education in Spanish of people responsible for documents at risk
3. Creation of the “Iberoamérica program” with the Secretaría General de Iberoamérica (Segib) and an IASA branch for preserving the Iberoamerican sound and audiovisual heritage using approved and shared standards.

**IASA Journal is constantly looking for material to publish:
articles, reviews, reports of meetings, or new developments.**

Please send articles, letters, or reviews you consider of interest to IASA, to the Editor at the address given at the back of this issue. Abstracts (maximum 250 words each) may be in French, German or English. Images can be sent as photographs or drawings to be scanned, or as digital images in GIF or TIFF formats.

The closing date for copy for the next issue, Number 39, to be published in June/July 2012, is 31 March 2012

Information for authors

Manuscript submission:

1. Articles should be submitted to the editor for consideration at least two months before published deadlines. If approved for publication, final versions of articles are due no later than the published deadlines.
2. Soft copy as a .doc file for text should be submitted with minimal formatting.
3. Illustrations (photographs, diagrams, tables, maps, etc) may be submitted as low resolution files placed in the .doc file or sent separately. Once the article has been accepted for publication, high resolution copies will be requested and should be sent as separate documents.
4. Use footnotes not endnotes and keep them to a minimum.
5. References should be listed at the end of the article in alphabetic order and chronologically for each author and should adhere to the following layout:

Christen, Kimberly. 2008. 'Archival Challenges and Digital Solutions in Aboriginal Australia.' *SAA Archaeological Recorder* 8(2):21-24.

Hyde, Lewis. 2010. *Common as Air: Revolution, Art, and Ownership*. New York: Farrar, Straus, and Giroux.

EVIA. 2010. EVIA Digital Archive Project: <http://www.eviada.org/> Consulted 23 October 2010.

6. Authors are encouraged to develop their conference presentations into more detailed accounts and/or arguments for publication in the journal. In principle, articles should be no longer than 5,000 words.

Information for advertisers

Enquiries about advertising should be sent to the Editor. Current rates can be seen on the website at <http://www.iasa-web.org/iasa-journal-advertising>

Join the IASA Listserv!

- Send a message **Subscribe lasalist@nb.no** + your **Full Name** (eg **Subscribe lasalist@nb.no name surname**) to: Listserv@nb.no
- Write only **Subscribe lasalist@nb.no + your Full Name** in the **text body**.
- Don't add anything else. **Leave the subject field empty**
- You will receive a message to confirm your subscription. Just follow the instructions.
- After a few seconds, you will receive a welcome message beginning:
- *You have been added to the IASALIST mailing list (IASA-list) by Meg administrator@NB.NO....*
- Please save this message for future reference, especially if this is the first time you are subscribing.
- You are subscribed and can start sending messages, questions, answers, etc to the list-serv.

IASA's sustaining members



<http://www.cube-tec.com/>



<http://www.noa-audio.com/>



<http://www.audioinspector.com/>



<http://www.av-inspector.com/>

IASA Journal

- Journal of the International Association of Sound and Audiovisual Archives IASA
- Journal de l'Association Internationale d'Archives Sonores et Audiovisuelles IASA
- Zeitschrift der Internationalen Vereinigung der Schall- und audiovisuellen Archive IASA
- El Journal de Asociación Internacional de Archivos Sonoros y Audiovisuales

The IASA Journal is published twice a year and sent to all the members of IASA. Applications for membership of IASA should be sent to the Secretary General (see list of officers below). The annual dues are €48 for individual members and €190 for institutional members. Back copies of the IASA Journal from 1971 are available on application. Subscription to the current year's issues of the IASA Journal is also available to non-members at a cost of €70.

Le IASA Journal est publié deux fois par an et distribué à tous les membres de l'association. Veuillez envoyer vos demandes d'adhésion au secrétaire don't vous trouverez l'adresse ci-dessous. Les cotisations annuelles se montent actuellement à €48 pour les membres individuels et €190 pour les membres institutionnels. Les anciens numéros (à partir de 1971) du IASA Journal sont disponibles sur demande. Ceux qui ne sont pas membres de l'Association peuvent s'abonner au IASA Journal pour l'année en cours au coût de €70.

Das IASA Journal erscheint zweimal jährlich und geht allen Mitgliedern der IASA zu. Aufnahmeanträge für die Mitgliedschaft bei der IASA sind an das Sekretariat (Anschrift siehe unten) zu richten. Die Mitgliedsbeiträge betragen derzeit €48 für individuelle Mitglieder und €190 für Institutionen. Frühere IASA Journale (ab 1971) sind auf Anfrage erhältlich. Der Bezugspreis des IASA Journals für Nichtmitglieder beträgt €70.

El 'IASA Journal' se publica dos veces al año y se envía a todos los miembros de la IASA. Las solicitudes de inscripción a la IASA deben dirigirse al Secretario General (consultar la lista de los miembros directivos a continuación). Las cuotas anuales son de €48 para los miembros individuales y de €190 para los institucionales. Los números atrasados del 'IASA Journal' desde 1971 están disponibles previa solicitud. Las suscripciones a los números del 'IASA Journal' del año en curso también están disponibles para no asociados, al precio de €70.

The Executive Board of IASA

President: **Jacqueline von Arb** Norwegian Institute of Recorded Sound, Norway. Email: jacqueline.vonarb@uis.no

Vice-Presidents: **Pio Pellizzari (Training)** Swiss National Sound Archives. Email: training@iasa-web.org
Alvaro Hegewisch (Membership) Fonoteca Nacional Mexico. Email: membership@iasa-web.org
Bruce Gordon (Conferences) Eda Kuhn Loeb Music Library, Music Building, North Harvard Yard, Harvard University. Email: conferences@iasa-web.org

Past President: **Kevin Bradley** Curator, Oral History and Folklore Director, Sound Preservation, National Library of Australia. Email: kbradley@nla.gov.au

Editor: **Cassandra Gallegos**. Email: cjg225@gmail.com

Web editor **Richard Ranft**. The British Library. Email richard.ranft@bl.uk

Secretary General **Lynn Johnson** e.tv Pty Limited, Cape Town, South Africa. Email: secretary-general@iasa-web.org

Treasurer: **Marit Hamre** NRK, Norway. Email: treasurer@iasa-web.org

© **The International Association of Sound and Audiovisual Archives IASA**
<http://www.iasa-web.org>

Authors grant an exclusive license to IASA to publish and disseminate to the widest possible readership in print and electronic formats online and via intermediaries as appropriate. Authors retain rights to use all or part of their article/s provided that a full acknowledgement is made to the original publication in the journal and that it is not published commercially. No part of this issue may be reproduced by any third party in any form or for any purpose without written permission from the publisher.

No part of this issue may be reproduced in any form, by print, photoprint, microfilm, or any other means without written permission from the publisher.

Signed articles and reviews represent the opinions of the authors and do not necessarily reflect the policies of the Association.

Layout by Smallgoodthing, London, UK. Printed by Bellsway Print, Orpington, UK.

IASA uses Gill Sans as its preferred font. **Gill Sans** was created by Eric Gill and published by the Monotype Corporation between 1928 and 1930. Gill Sans is widely admired for its quiet gracefulness and versatility. In the Bit stream font collection, Gill Sans is called Bit stream Humanist 521. Gill Sans was part of a competitive period in the 1920s when various foundries were developing modern sans-serif type faces for various lead type setting technology.

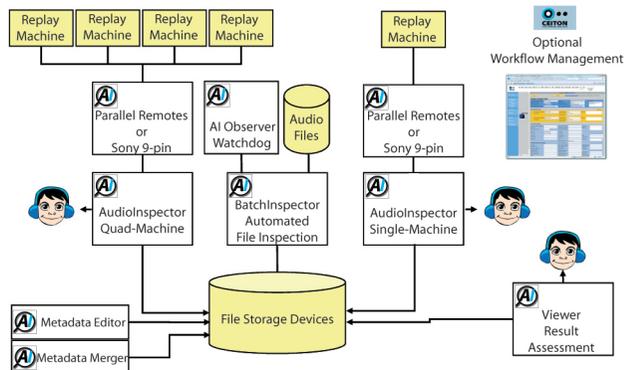
EBU R128 loudness analysis for archive content

AUDIOINSPECTOR analyses & evaluates:

- Loudness according EBU R128 incl. True Peak
- Duration of the recording
- Amount of segments & breaks in the recording
- Maximum silence duration
- Amount of stereo and mono segments
- Amount of speech and music segments
- Stereo balance
- Bandwidth incl. upper and lower border frequency
- Maximum DC-offset
- Average digital level, digital peak & digital silence level
- Signal to noise ratio
- Digital over levels
- Analog clipping
- Azimuth
- Correlation
- Clicks
- Pops
- Drop-outs
- Digital error concealments
- Cut beginnings and endings

- Parallel digitization and automated file analysis
- Optional workflow management tool
- Interfaces to restoration tools
- Flat fee or Pay-per-Use licensing model
- Running on Windows 7 (32 & 64 Bit), Vista & XP

The AudioInspector Product Family



"We are impressed by the technology of AudioInspector and the possibility to integrate the new EBU core metadata set into our workflow.

We are using AudioInspector for transferring titles out of our 250.000 DAT-tape collection, with other formats to follow."

Thomas Bårdsen
Production Manager
The National Library of Norway



Try before you buy!

**Download your fully functional AudioInspector
for free testing at www.audioinspector.com**

*Of course we have an **archive!****



* Is the most common argument we hear before clients have discovered our archive solutions.

mediARC™

Metadata – media – workflows.
The flexible media archive.



jobDB™

This is not an archive.
A subsystem for QC and Media
workflows in your archive.



actLINE™

Processors to reshape your
content.



ingestLINE™

Processors for the transcription
of archive carriers.



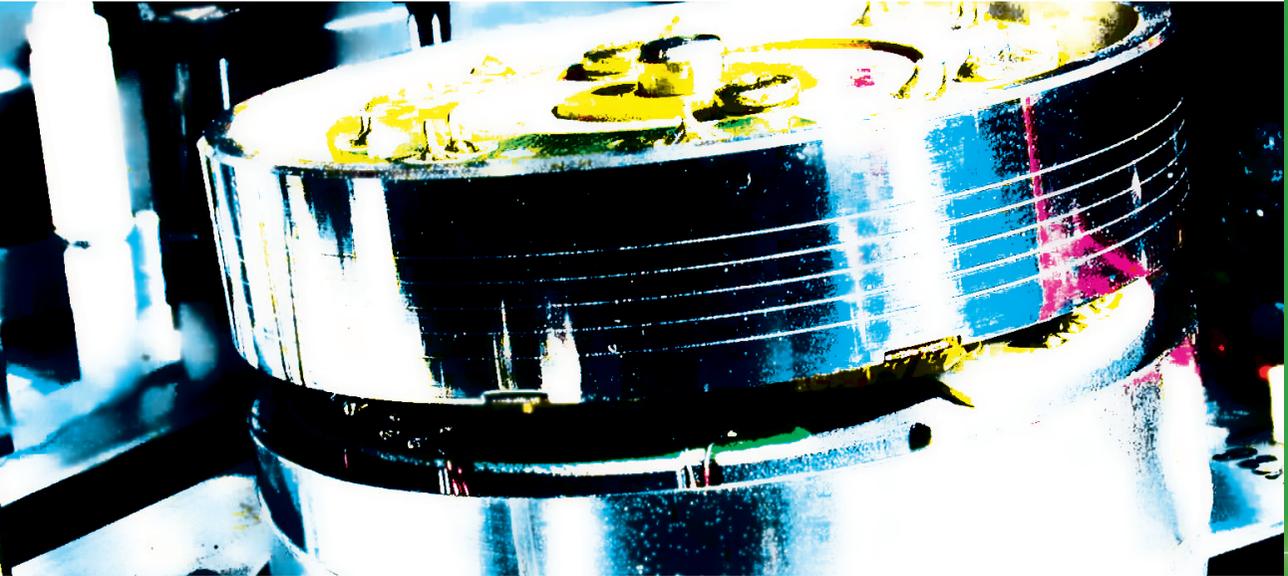
NEW



QUADRIGA

VIDEO

Quality-Controlled Mass-Digitization for Video



- **Legacy Media Ingest**
- **Process-Automation**
- **Workflow-Optimization**



www.Cube-Tec.com



Cube-Tec supports IASA - we are a sustaining member for more than ten years.