CHANGING SCENERIES
CHANGING ROLES PART IV
KEEPING YOUR BEST CONTENT AND METADATA

SELECTED PAPERS FROM THE FIAT/IFTA MEDIA MANAGEMENT SEMINAR, STOCKHOLM 2009
CHANGING SCENERIES
CHANGING ROLES PART IV
KEEPING YOUR BEST CONTENT AND METADATA

PROCEEDINGS OF THE MEDIA MANAGEMENT SEMINAR
CHANGING SCENERIES, CHANGING ROLES PART IV
STOCKHOLM 15TH - 16TH OF MAY 2009

ORGANISED BY
THE FIAT/IFTA MEDIA MANAGEMENT COMMISSION
THE SWEDISH NATIONAL LIBRARY AND
SVT - SVERIGES TELEVISION

DESIGNED BY AXEL GREEN FOR FIAT/IFTA
ORIGINALLY PRINTED BY BEELD EN GELUID AT DRUKKERIJ DE TOEKOMST
CONTENTS

PREFACE  
   by Eva-lis Green  
   4

PRINCIPALS OF SELECTION  
   by Herbert Hayduck  
   6

1. SELECTION REVISITED  
   by Annemiek de Jong  
   8

2. AN INTRODUCTION TO CULTURAL ANALYTICS  
   by Lev Manovich  
   16

3. SELECTION IN THE DIGITAL DOMAIN  
   by Marie-Anne Chabin  
   44

4. MEMENTO-PROJECT  
   by Robbie de Sutter  
   62

5. METADATA QUALITY CRITERIA  
   by Anna Abenius  
   70

6. RETENTION POLICY AT BBC SCOTLAND  
   by Vicky Plaine  
   84

7. YLE CENTRALIZED INGEST AREA  
   by Henry Lindqvist  
   98
8. SUSTAINABILITY & INFRASTRUCTURE
   by Gerhard Stanz

9a. SAVE AND DIGITIZE ALL ARCHIVE?
   by Dominique Saintville

9b. LA STRATÉGIE DE L’INA, 1999-2015
   by Dominique Saintville

10. MANAGING ABUNDANCE AND DIVERSITY
    by Claude Mussou

11. SELECTION AND RETENTION
    by Lisette Graswinckel & Irma van Kampen

12. PRESERVING THE BEST METADATA AND CONTENT
    by Göran Konstenius

14. CHANGING ROLES IN A TAPELESS ENVIRONMENT
    by Eric Minoli

OBSERVATIONS
    by Jacqui Gupta

INDEX OF STILLS
ON THE 14-15th May 2009 a MMC Seminar was held in the National Library in Stockholm titled Changing Sceneries, Changing Roles – part IV. The seminar theme was: Keeping your best content & metadata, Retention Management in the Digital Domain. The objective was to answer the questions in this area by combining theory and practical projects. Some of the presentations are collated in this book.

The seminar, organized by the Media Management Commission (FIAT/IFTA) in cooperation with Sweden National Library and Sveriges Television, addressed the perspectives and interests of both broadcasters and national archives. The overall focus was to try to conclude the new challenges to the archivist role, the archive business and archival thinking.

The last chapter of the book is entitled observations, not conclusions. The reason is that the discussion must go on, even after this book, as there are no simple solutions to the issues.

The head of the National Library Gunnar Sahlin and the CEO from Sveriges Television Eva Hamilton gave the seminar’s first contributions
with their welcome speeches. Gunnar Sahlin welcomed the delegates to the National Library and also pointed out the convergence between different types of media in the new information landscape. Eva Hamilton introduced the audience to the MAM-system used in Sveriges Television. She stressed the values of archives but also the big challenges concerning storage and access.

The meeting was opened by FIAT/IFTA president Herbert Hayduck. The first chapter of this book is an illustration of Herbert Hayduck’s presentation about the principles for Selection. It gives a good overview to the subject.

Two well known experts from different domains other than Audiovisual Archives were also invited; Marie-Anne Chabin from the traditional archive domain and Lev Manovich from the academic, new media research sphere. Their presentations are included in this book and gives important perspectives and input from other areas in the digital domain regarding the issue of selection.

The seminar was moderated by Sissela Andrén from Sveriges Television. Along with speeches and presentations a lively audience activity was performed and also panel debates. Some of the questions asked were: What will you keep as a private person, as a professional and what do you want our society to keep? These simple questions made at least me and hopefully also the other delegates to reflect on what to keep in a new perspective.

I want to thank the National Library, Sveriges Television staff involved, the organizing Committee Göran Konstenius, Jacqui Gupta, Annemieke de Jong and Pelle Snickars and all the speakers and contributors to this publication. Thanks also to SVT Bild for letting us use the images, that apart from being archive material themselves, makes the book even more enjoyable. Thank you all!

Eva-Lis Green
Chairman, MMC Commission
Stockholm, september 2010
### Key Questions for Selection
- What for?
- For Whom?
- How?
- What?
- When?
- Who?

### Principles of Preservation
- Save the content, not the carrier
- First things first
  - Priorisation and Selection
  - Influenced by mission statement of archive institution
- Target of preservation are future audiences, not the total rescue of all existing material
- The future value of material defines
  - the amount of money spent for storage and access technologies
  - the amount of effort spent for transfer and quality control

### Content
- Integral Re-Broadcast or partial Reuse
- Broadcast Programme or Rushes
- Content Genres:
  - News, sports, entertainment etc.
  - Representative Documentation of
  - Broadcast History
  - Historical Value
  - Representation of Contemporary History

### Copyright
- Technical Urgency for Transfer
- Obsolescence of
  - Carriers
  - Equipment
- Priorisation of
  - Carriers
  - Equipment
- Individual action or joint solution necessary

### Format
- Type of Production
  - Own
  - External
- Character of Reuse
  - Commercial
  - Non-commercial
- Type of Reuse
  - Integral
  - Partial
Annemieke de Jong is a long standing member of the Media Management Commission of FIAT and member of the Seminar Programme Committee. She works at the Netherlands Institute for Sound and Vision, the national audiovisual archive and central broadcast archive of the Netherlands. As corporate policy advisor she is responsible for the development and dissemination of knowledge on the theoretical and practical aspects of (digital) audiovisual archiving. On a national and an international level Annemieke de Jong is involved in several projects and expert groups within the field of media management and digital heritage. Annemieke de Jong holds MA degrees in New Media and Digital Culture (University of Utrecht) and in Information Management (Erasmus University Rotterdam).
IN THIS 4TH edition of the seminar “Changing Sceneries, Changing Roles” we once again explore aspects of the new digital space and what it means to us as audiovisual archivists. During our very first edition, in London at the end of the nineties, we focused on the emerging digital production environment, trying to understand what we saw coming. Part two took place in Amsterdam in 2004. There we discussed the developing media systems in a tapeless production setting. We also conceptualized the new archive professional: the media manager. Our third 2007 seminar in Vienna was dedicated to the integrated digital workflows that were by that time actually in place. We also discussed the developments in metadata creation.

The fourth Stockholm edition deals with what is perhaps one of the core issues within digital production and archiving: selection and retention of the actual content and metadata.
The situation can be described as follows. In the professional production environment much more material – be it items, clips, rushes, websites, un-edited materials and final products – is being created than ever before. This explosive and heterogeneous production is partly caused by the increase in publications means, such as new (web) platforms, intended to service new online users and support new types of usage, other than the professional. It’s also highly stimulated by digital production systems, linked in an integrated workflow, that facilitate and speed up continuous creation, publication and circulation of content and metadata. This tremendous increase in volume, dynamics and publication creates new dilemmas, but also new perspectives on building and preserving archival collections. These developments force us to rethink the basic questions surrounding selection in the digital domain.

**SELECTION: WHY?**

To start with: why should selection be an issue in the first place? Why select at all?

We’re all aware that selection takes up a lot of precious time and effort while at the same time digital storage is cheap and getting cheaper every day. With digital content it all comes down to providing access anyway and we know that it will no longer be necessary to index everything manually. A lot of serious help is arising out of developments in advanced search technology, research on automated video, audio and text analysis, the rise of semantic technologies and last but not least: the emergence of social tagging and other forms of user generated metadata (UGM).

Selection is always arbitrary. In not selecting, in keeping everything, we would – as audiovisual archivists – for the first time be in line with the ideas of great archival thinkers like Sir Hilary Jenkinson and his colleagues. They felt that archives are principally there to keep things, not to
throw things out. Archives should function as neutral, impartial institutes and any selection undergone during acquisition is wrong. The archivist should not play God, he is a keeper.

Obviously Sir Hilary could not anticipate the tidal wave that would come to flood us, the endless stream of edited and unedited digital content that is copied, intertwined, versioned or all three at the same time. Will we be doing our current and future users a favor by keeping all this? And are we really so sure that storage is not a problem any more? The alarming issues here around HD illustrate that there is still a lot to be solved in that area. Providing access to everything that is being produced – even if its done largely automatically – will still need a lot of human effort, from providing input to research programmes on automated indexing to the management of UGM around institutional collections. Finally, we should keep in mind that everything we decide to retain and preserve requires maintenance. It will at one point have to be migrated to a new digital format and another technological environment. Responsibility for sustainable preservation in order to keep material accessible will always continue!

It seems that – if we at least want to offer some quality and balance in what we preserve for posterity – there is a definite need to control the tidal wave, for more than one reason. Setting up selection and retention policies is clearly one way to fulfill that need.

THE WHEN ISSUE

In the digital production environment, editing, distribution, publication and storage is ongoing. Digital content and metadata grows in organic ways and bits and pieces of both are used and re-used throughout the process as well as after. These media-objects become moving targets that are hard to keep track of. Archiving the material is no longer a well defined activity at the end of the chain. It’s done all over the place, either
automatically by the systems, or manually, by participants in production. The question is: when to rationally decide a digital object should be preserved temporarily or for the long term? At which point to intervene in the digital chain and assign the objects as either “heritage” or production material?

A complicating factor is that content value changes over time. In the analogue days – when collections could still be physically handled – archives were able to accept material and decide later whether or not to invest in it, by way of extensive cataloguing and/or active conservation for example. The decision to acquire was separated from the decision to preserve materials for the long term. And given this ability to decide over time, decision making could be based on appraisal selection criteria. In the digital environment, acquisition and preservation decisions seem to collide. The decision to retain and preserve material on costly sustainable formats must be taken very early in the production process because upwards compatibility in a later stage will not be possible. Two new problems arise here: the huge amounts of material to choose from and the lack of a certain time period allowing one time to reflect and appraise content for its lasting, heritage value.

**WHAT SHOULD BE KEPT?**

In the digital environment, archival collections are no longer built out of discrete entities, such as films and tapes neatly and promptly delivered to the archive after publication or transmission. Digital ‘collections’ are dynamic, distributed and undefined. It’s rather hard to say where they end and where they begin. Any part may be digitally born or digitized, and stem from many different sources, created by different creators, generated during different production stages. On top of this, new publication means create divergence: the same content element may be used in a television
programme, as part of a website, on a digital channel, on a DVD or in a videostream to a mobile device. These can be seen as different representations of the same source material.

The archives’ mission is twofold: to support the production process and to preserve a company’s heritage. Taking both these interests into account, what then should we choose to keep: the media-items themselves, their representations, the metadata or perhaps all?

It’s clear that when selecting for either of those purposes, it’s necessary to distinguish between different rationales. We have to outline what we want to keep in order to be able to support production and to preserve heritage value. Obviously this will include both source materials (the media-items themselves), as well as (some of) their representations and the associated metadata. Secondly we must be aware of what we need to keep, acknowledging the issue of interdependency that implies that some representations cannot be reconstructed without the associated digital objects and/or metadata. And finally we have to know precisely what we can keep. Some representations cannot be reconstructed at all because they have too many interactive elements. For these materials it is merely possible to keep the metadata that describes them.

**HOW TO SELECT**

The what and when questions surrounding selection in the digital production domain cannot be separated from the “how”. On the practical level we need to learn how to integrate long term interests in a highly dynamic workflow. We have to decide on the right moments to intervene in the production chain, by defining points for a virtual “handover” of materials to be retained for either the short or the long term. For this we should know how to keep track of materials throughout their creation, production and distribution. We need to start adequately managing the lifecycles
of digital materials, by installing fixed protocols for their identification, authentication and transition. Digital technology itself may very well support this, helping us better cope with the volume, versions and dynamics. MAM systems can be equipped with useful functionality, that automates and standardizes large parts of the selection process, through naming conventions and statistical functions for example. Finally, on a more organizational level, we have to engage stakeholders in the selection and retention process. Now that the archive is no longer solely responsible for what is being kept and what is being thrown out, it should be producers, programmers, editors, technicians and other users that must be made aware of the importance of standardized selection protocols so that they can help design them and carry them out.

**CHANGING SCENARIOS, CHANGING SELECTION**

The growing volume and the new dynamics of professional digital production are starting to deeply effect our collections, our workflows and our roles as archivists, be it in the context of the broadcast process we support, or wider, as keepers of media heritage. Rethinking the basic questions surrounding selection, and the complicated way these questions hang together in the digital domain, is a way to envision these effects and – to a certain extent – control them. The programme committee will have reached its goal when the ideas, projects and conversations presented during this seminar contribute to this process.

*Annemieke de Jong*

*Seminar Programme Committee*
Lev Manovich is a Professor in Visual Arts Department, University of California, San Diego, a Director of the Software Studies Initiative at California Institute for Telecommunications and Information Technology (Calit2), and a Visiting Research Professor at Goldsmith College (University of London), De Montfort University (UK) and College of Fine Arts, University of New South Wales (Sydney). He is much in demand to lecture around the world, having delivered over 300 lectures, seminars and workshops during the last ten years. Lev Manovich’s books include *Software Takes Command* (released under CC license, 2008), *Soft Cinema: Navigating the Database* (The MIT Press, 2005), and *The Language of New Media* (The MIT Press, 2001) which is hailed as “the most suggestive and broad ranging media history since Marshall McLuhan.” He has written over 90 articles which have been reprinted over 300 times in over 30 countries.
ONLy FIFTEEn yEARS ago we typically interacted with relatively small bodies of information that were tightly organized in directories, lists and a priori assigned categories. Today we interact with a gigantic, global, not well organized, constantly expanding and changing information cloud in a very different way: we Google it.

The raise of search as the new dominant way for encountering information is one manifestation of the fundamental change in human’s information environment. We are living through an exponential explosion in the amounts of data we are generating, capturing, analyzing, visualizing, and storing – including cultural content. On August 25, 2008, Google’s software engineers announced on googleblog.blogspot.com that the index of web pages, which Google is computing several times daily, has reached 1 trillion unique URLs. During the same month, YouTube.com reported that users were uploaded 13 hours of new video to the site every minute.
and in November 2008, the number of images housed on Flickr reached three billions.

The “information bomb” already described by Paul Virilio in 1998 has not only exploded. It also led to a chain of new explosions that together produced cumulative effects larger than anybody could have anticipated. In 2008 International Data Corporation (IDC) forecasted that by 2011, the digital universe would be ten times the size it was in 2006. This corresponds to a compound annual growth rate of 60 percent. Of course, it is possible that the global economic crisis which begun in 2008 may slow this growth – but probably not too much.

User-generated content is one of the fastest growing parts of this expanding information universe. According to IDC 2008 study, “Approximately 70 percent of the digital universe is created by individuals.” In other words, the size of media created by users competes well with the amounts of data collected and created by computer systems (surveillance systems, sensor-based applications, datacenters supporting “cloud computing,” etc.) So if Friedrich Kittler – writing well before the phenomena is “social media” – noted that in a computer universe “literature” (i.e. texts of any kind) consists mostly of computer-generated files, the humans are now catching up.

The exponential growth of a number of both non-professional media producers in the 2000’s has led to a fundamentally new cultural situation and a challenge to our normal ways of tracking and studying culture. Hundreds of millions of people are routinely creating and sharing cultural content – blogs, photos, videos, map layers, software code, etc. The same hundreds of millions of people engage in online discussions, leave comments and participate in other forms of online social communication. As the number of mobile phones with rich media capabilities is projected to keep growing, this number is only going to increase. In early 2008, there were 2.2 billion mobile phones in the world; it was projected that this number will become 4 billion by 2010, with main growth coming from China, India, and Africa.
Think about this: the number of images uploaded to Flickr every week today is probably larger than all objects contained in all art museums in the world.

The exponential increase in the numbers of non-professional producers of cultural content has been paralleled by another development that has not been widely discussed. And yet this development is equally important in understanding what culture is today. The rapid growth of professional educational and cultural institutions in many newly globalize countries since the end of the 1990’s – along with the instant availability of cultural news over the web and ubiquity of media and design software – has also dramatically increased the number of culture professionals who participate in global cultural production and discussions. Hundreds of thousands of students, artists, designers, musicians have now access to the same ideas, information and tools. As a result, often it is no longer possible to talk about centers and provinces. In fact, based on my own experiences, I believe the students, culture professionals, and governments in newly globalized countries are often more ready to embrace latest ideas than their equivalents in “old centers” of world culture.

If you want to see the effects of these dimensions of cultural and digital globalization in action, visit the popular web sites where the professionals and the students working in different areas of media and design upload their portfolios and samples of their work – and note the range of countries from which the authors come from. Here are examples of these sites: xplsv.tv (motion graphics, animation), coroflot.com (design portfolios from around the world), archinect.com (architecture students projects), infosthetics.com (information visualization projects). For example, when I checked on December 24, 2008, the first three projects in the “artists” list on xplsv.tv came from Cuba, Hungary, and Norway. Similarly, on the same day, the set of entries on the first page of coroflot.com (the site where designers from around the world upload their portfolios; it contained over 120,000 portfolios by the beginning of 2009) revealed a similar global
cultural geography. Next to the predictable 20th century Western cultural capitals – New York and Milan – I also found portfolios from Shanghai, Waterloo (Belgium), Bratislava (Slovakia), and Seoul (South Korea).

The companies which manage these sites for professional content usually do not publish detailed statistics about their visitors – but here is another example based on the quantitative data which I do have access to. In the spring of 2008 we have created a web site for our research lab at University of California, San Diego: softwarestudies.com. The web site content follows the genre of “research lab site” so we did not expect many visitors; we also have not done any mass email promotions or other marketing. However, when I examined Google Analytics stats for softwarestudies.com at the end of 2008, I discovered that we had visitors from 100 countries. Every month people from over 1000 cities worldwide check out the site. Even more interestingly are the statistics for these cities. During a typical month, no American cities made it into “top ten list” (I am not counting La Jolla which is the location of UCSD where our lab is located). For example, in November 2008, New York occupied 13th place, San Francisco was at 27th place, and Los Angeles was at 42nd place. The “top ten” cities were from Western Europe (Amsterdam, Berlin, Porto), Eastern Europe (Budapest), and South America (Sao Paulo). What is equally interesting is the list of visitors per city followed a classical “long tail” curve. There was no sharp break anymore between “old world” and “new world,” or between “centers” and “provinces.” (See softwarestudies.com/softbook for more complete statistics.)

All these explosions which took place since the late 1990s – non-professionals creating and sharing online cultural content, culture professionals in newly globalized countries, students in Eastern Europe, Asia and South America who can follow and participate in global cultural processes via the web and free communication tools (email, Skype, etc) – redefined what culture is.
Before, cultural theorists and historians could generate theories and histories based on small data sets (for instance, “classical Hollywood cinema,” “Italian Renaissance,” etc.) But how can we track “global digital cultures” with their billions of cultural objects, and hundreds of millions of contributors? Before you could write about culture by following what was going on in a small number of world capitals and schools. But how can we follow the developments in tens of thousands of cities and educational institutions?

INTRODUCING CULTURAL ANALYTICS

The ubiquity of computers, digital media software, consumer electronics, and computer networks led to the exponential rise in the numbers of cultural producers worldwide and the media they create – making it very difficult, if not impossible, to understand global cultural developments and dynamics in any substantial details using 20th century theoretical tools and methods. But what if we can use the same developments – computers, software, and availability of massive amounts of “born digital” cultural content – to track global cultural processes in ways impossible with traditional tools?

To investigate these questions – as well as to understand how the ubiquity of software tools for culture creation and sharing changes what “culture” is theoretically and practically – in 2007 we established Software Studies Initiative (softwarestudies.com). Our lab is located at the campus of University of California, San Diego (UCSD) and it is housed inside one of the largest IT research centers in the U.S. – California Institute for Telecommunications and Information (www.calit2.net). Together with the researchers and students working in our lab, we have been developing a new paradigm for the study, teaching and public presentation of cultural artifacts, dynamics, and flows. We call this paradigm Cultural Analytics.
Today sciences, business, governments and other agencies rely on computer-based quantitative analysis and interactive visualization of large data sets and data flows. They employ statistical data analysis, data mining, information visualization, scientific visualization, visual analytics, simulation and other computer-based techniques. Our goal is start systematically applying these techniques to the analysis of contemporary cultural data. The large data sets are already here – the result of the digitization efforts by museums, libraries, and companies over the last ten years (think of book scanning by Google and Amazon) and the explosive growth of newly available cultural content on the web.

We believe that a systematic use of large-scale computational analysis and interactive visualization of cultural patterns will become a major trend in cultural criticism and culture industries in the coming decades. What will happen when humanists start using interactive visualizations as a standard tool in their work, the way many scientists do already? If slides made possible art history, and if a movie projector and video recorder enabled film studies, what new cultural disciplines may emerge out of the use of interactive visualization and data analysis of large cultural data sets?

FROM CULTURE (FEW) TO CULTURAL DATA (MANY)

In April 2008, exactly one year later we founded Software Studies Initiative, NEH (National Endowment for Humanities, the main federal agency in the U.S. which provides grants for humanities research) announced a new “Humanities High-Performance Computing” (HHPC) initiative that is based on the similar insight:

“Just as the sciences have, over time, begun to tap the enormous potential of High-Performance Computing, the humanities are beginning to as well. Humanities scholars often deal with large sets of unstructured data. This might
take the form of historical newspapers, books, election data, archaeological fragments, audio or video contents, or a host of others. HHPC offers the humanist opportunities to sort through, mine, and better understand and visualize this data.

In describing the rationale for Humanities High-Performance Computing program, the officers at NEH start with the availability of high-performance computers that are already common in the sciences and industry. In January 2009, NEH together with NSF (National Science Foundation) has announced another program Digging Into Data which has articulated their vision in more detail. This time the program statement put more emphasis on the wide availability of cultural content (both contemporary and historical) in digital form as the reason for begin applying data analysis and visualization to “cultural data.”

With books, newspapers, journals, films, artworks, and sound recordings being digitized on a massive scale, it is possible to apply data analysis techniques to large collections of diverse cultural heritage resources as well as scientific data. How might these techniques help scholars use these materials to ask new questions about and gain new insights into our world?

We fully share the vision put forward by NEH Digital Humanities. Massive amounts of cultural content and high-speed computers go well together – without the latter, it would be very time consuming to analyze petabytes of data. However, as we discovered in our lab, even with small cultural data sets consisting from hundreds, dozens or even only a few objects it is already viable to do Cultural Analytics: that is, to quantitatively analyze the structure of these objects and visualize the results revealing the patterns which lie below the unaided capacities of human perception and cognition.

Since Cultural Analytics aims to take advantage of the exponential increase in the amounts of digital content since the middle of the 1990’s, it will be useful to establish taxonomy for the different types of this content.
Such taxonomy may guide design of research studies as well as be used to group these studies once they start multiply.

To begin with, we have vast amounts of media content in digital form—games, visual design, music, video, photos, visual art, blogs, web pages. This content can be further broken down into a few categories. Currently, the proportion of “born digital” media is increasing; however, people also continue to create analog media (for instance, when they shoot on film), which is later digitized.

We can further differentiate between different types of “born digital” media. Some of this media is explicitly made for the web: for example, blogs, web sites, layers created by users for Google Earth and Google maps. But we also now find online massive amounts of “born digital” content (photography, video, music) which until the advent of “social media” was not intended to be seen by people worldwide—but which now ends up online at social media sites (Flickr, YouTube, etc.) To differentiate between these two types, we may refer to the first category as “web native,” or “web intended.” The second category can be then called “digital media proper.”

As I already noted, YouTube, Flickr, and other social media sites aimed at average people are paralleled by more specialized sites which serve professional and semi-professional users: xplsv.tv, coroflot.com, archinect.com, modelmayhem.com, deviantart.com, etc. Housing projects and portfolios by hundreds of thousands of artists, media designers, and other cultural professionals, these web sites provide a live snapshot of contemporary global cultural production and sensibility—thus offering a promise of being able to analyze the global cultural trends with the level of detail unthinkable previously. For instance, as of August 2008, deviantart.com has eight million members, over 62 million submissions, and was receiving 80,000 submissions per day. Importantly, in addition to the standard “professional” and “pro-ams” categories, these sites also house the content of people who are just starting out and/or are currently “pro-ams” but who aspire to be full-time professionals. I think that the portfolios (or “ports” as
they are sometimes called today) of these “aspirational non-professionals” are particularly significant if we want to study contemporary cultural stereotypes and conventions since, in aiming to create “professional” projects and portfolios, people often inadvertently expose the codes and the templates used in the industry in a very clear way.

Another important source of contemporary cultural content – and at the same time, a window into yet another cultural world different from non-professional users and aspiring professionals – are the web sites and wikis created by faculty teaching in creative disciplines to post and discuss their class assignments. Although I don’t have direct statistics on how many sites and wikis for classes are out there, here is one indication: a popular wiki creation software pbwiki.com has been used by 250,000 educators. These sites often contain student projects – which provides yet another interesting source of content.

Finally, beyond class web sites, the sites for professionals, aspiring professionals, and non-professionals, and other centralized content repositories, we have millions of web sites and blogs by individual cultural creators and creative industry companies. Regardless of the industry category and the type of content people and companies produce, it is now taken for granted that you need to have a web presence with your demo reel and/or portfolio, descriptions of particular projects, a CV, and so on. All this information can be potentially used to do something that previously was un-imaginable: to create dynamic (i.e. changing in time) maps of global cultural developments that reflect activities, aspirations, and cultural preferences of millions of creators.

A significant part of the available media content in digital form was originally created in electronic or physical media and has been digitized since the middle of the 1990’s. We can call such content “born analog.” But it is crucial to remember that what has been digitized in many cases are only the canonical works, i.e. a tiny part of culture deemed to be significant
by our cultural institutions. What remains outside of the digital universe is the rest: provincial nineteen century newspapers sitting in some small library somewhere; millions of paintings in tens of thousands of small museums in small cities around the world; millions of thousands of specialized magazines in all kinds of fields and areas which no longer even exist; millions of home moves…

This creates a problem for Cultural Analytics, which has a potential to map everything that remains outside the canon – to begin generating “art history without great names.” We want to understand not only the exceptional but also the typical; not only the few “cultural sentences spoken by a few “great man” but the patterns in all cultural sentences spoken by everybody else; in short, what is outside a few great museums rather than what is inside and what has been already extensively discussed too many times. To do this, we will need as much of previous culture in digital form as possible. However, what is digitally available is surprisingly little.

Here is an example from our research. We were interested in the following question: what did people actually paint around the world in 1930 – outside of a few “isms” and a few dozen artists who entered the Western art historical canon? We did a search on artstor.org which at the time of this writing contains close to one million images of art, architecture and design which come from many important US museum and collections, as well as over 200,000 slides library of University of California, San Diego where our lab is located. This set which at present is the largest single collection in artstor is interesting in that it reflects the biases of art history as it was taught over a few decades when color slides were the main media for teaching and studying art. To collect the images of artworks that are outside of the usual Western art historical canon, we excluded from the search Western Europe and North America. This left the rest of the world: Eastern Europe, South-East Asia, East Asia, West Asia, Oceania, Central America, South America, etc. When we searched for paintings done in these parts of the world in 1930, we only found a few dozen images. This
highly uneven distribution of cultural samples is not due to Artstor since it does not digitize images itself – it only makes available images submitted to its by museums and other cultural institutions. So what the results of our search reflect is what museums collect and what they think should be digitized first. In other words, a number of major US collections and a slide library of a major research university (which now has a large proportion of Asian students) together contain only a few dozen paintings done outside the West in 1930 which got digitized. In contrast, searching for Picasso returned around 700 images. If this example if any indication, digital depositories may be amplifying the already existed biases and filters of modern cultural canons. Instead of transforming the “top forty” into “the long tail,” digitization can be producing the opposite effect.

Media content in digital form is not the only type of data that we can analyze quantitatively to potentially reveal new cultural patterns. Computers also allow us to capture and subsequently analyze many dimensions of human cultural activities that could not be recorded before. Any cultural activity – surfing the web, playing a game, etc. – which passes through a computer or a computer-based media device leaves traces: keystroke presses, cursor movements and other screen activity, controller positions (think of We controller), and so on. Combined with camera, a microphone, and other capture technologies, computers can also capture other dimensions of human behavior such as body and eye movements and speech. And web servers log yet other types of information: which pages the users visited, how much time they spend on each page, which files they downloaded, and so on. (In this respect, Google Analytics that processes and organizes this information provided a direct inspiration for the idea of Cultural Analytics.

Of course, in addition to all this information which can be captured automatically, the rise of social media since 2005 created a new social environment where people voluntarily reveal their cultural choices and
preferences: rating books, movies, blog posts, software, voting for their favorites, etc. Even importantly, people discuss and debate their cultural preferences, ideas and perceptions online. They comment on Flickr photographs, post their opinions about books on amazon.com, critique movies on rottentomatoes.com, review products on epinions.com, and enthusiastically debate, argue, agree and disagree with each other on numerous social media sites, fan sites, forums, groups, and mailing lists. All these conversations, discussions and reflections which before were either invisible or simply could not take place on the same scale are now taking place in public.

To summarize this discussion: because of digitization efforts since the middle of the 1990s, and because the significant (and constantly growing) percentage of all cultural and social activities passes through, or takes place on the web or networked media devices (mobile phones, game platforms, etc.), we now have access unprecedented amounts of both “cultural data” (cultural artifacts themselves), and “data about culture.” All this data can be grouped into three broad conceptual categories:

- Cultural artifacts (“born digital” or digitized).
- Data about people’ interactions with digital media (automatically captured by computers or computer-based media devices)
- Online discourse around (or accompanying) cultural activities, cultural objects, and creation process voluntarily created by people.

There are other ways to divide this recently emerged cultural data universe. For example, we can also make a distinction between “cultural data” and “cultural information”: 
• Cultural data: photos, art, music, design, architecture, films, motion graphics, games, web sites – i.e., actual cultural artifacts which are either born digital, or are represented through digital media (for examples, photos of architecture).

• Cultural information: cultural news and reviews published on the web (web sites, blogs) – i.e., a kind of “extended metadata” about these artifacts.

Another important distinction, which is useful to establish, has to do with the relationships between the original cultural artifact/activity and its digital representation:

• “Born digital” artifacts: representation = original.

• Digitized artifacts that originated in other media – therefore, their representation in digital form may not contain all the original information. For example, digital images of paintings available in online repositories and museum databases normally do not fully show their 3D texture. (This information can be captured with 3D scanning technologies – but this is not commonly done at this moment.)

• Cultural experiences (experiencing theatre, dance, performance, architecture and space design; interacting with products; playing video games; interacting with locative media applications on a GPS enabled mobile device) where the properties of material/media objects that we can record and analyze is only one part of an experience. For example, in the case of spatial experiences, architectural plans will only tell us a part of a story; we may also want to use video and motion capture of people interacting with the spaces, and other information.
The rapid explosion of “born digital” data has not passed unnoticed. In fact, the web companies themselves have played an important role in making it happen so they can benefit from it economically. Not surprisingly, out of the different categories of cultural data, born digital data is already been exploited most aggressively (because it is the easiest to access and collect), followed by digitized content. Google and other search engines analyze billions of web pages and the links between them to make their search algorithms run. Nielsen Blogpulse mines over 100 million blogs to detect trends in what people are saying about particular brands, products and other topics its clients are interested in. Amazon.com analyzes the contents of the books it sells to calculate “Statistically Improbable Phrases” used to identify unique parts of the books.

In terms of media types, today text receives most attention – because language is discrete and because the theoretical paradigms to describe it (linguistics, computational linguistics, discourse analysis, etc.) have already been fully developed before the explosion of “web native” text universe. Another type of cultural media, which is also starting to be systematically subjected to computer analysis in large quantities, is music. (This is also made possible by the fact that Western music used formal notation systems for a very long time.) A number of online music search engines and Internet radio stations use computation analysis to find particular songs. (Examples: Musipedia, Shazam, and other applications which use acoustic fingerprinting.) In comparison, other types of media and content receive much less attention.

If we are interested in analyzing cultural patterns in other media besides text and sound, and also in asking larger theoretical questions about cultures (as opposed to more narrow pragmatic questions asked in professional fields such as web mining or quantitative marketing research – for instance, identifying how consumers perceive different brands in a particular market segment), we need to adopt a broader perspective. Firstly, we need to develop techniques to analyze and visualize the patterns in
different forms of cultural media – movies, cartoons, motion graphics, photography, video games, web sites, product and graphic design, architecture, etc. Second, while we can certainly take advantage of the “web native” cultural content, we should also work with other categories that I listed above ("digitized artifacts which originated in other media"); “cultural experiences.”) Thirdly, we should be self-reflective. We need to think about the consequences of thinking of culture as data and of computers as the analytical tools: what is left outside, what types of analysis and questions get privileged, and so on. This self-reflection should be part of any Cultural Analytics study. These three points guide our Cultural Analytics research.

**CULTURAL IMAGE PROCESSING**

Cultural Analytics is thinkable and possible because of three developments: digitization of cultural assets and the rise of web and social media; work in computer science; and the rise of a number of fields which use computers to create new ways of representing and interacting with data. The two related fields of computer science – image processing and computer vision – provide us with the variety of techniques to automatically analyze visual media. The fields of science visualization, information visualization, media design, and digital art provide us with the techniques to visually represent patterns in data and interactively explore this data.

While people in digital humanities have been using statistical techniques to explore patterns in literary text for a long time, I believe that we are the first lab to start systematically using image processing and computer vision for automatic analysis of visual media in the humanities contest. This is what separates us from 20th century humanities disciplines that focus on visual media (art history, film studies, cultural studies) and also 20th century paradigms for quantitative media research developed within social
sciences such as quantitative communication studies and certain works in sociology of culture. Similarly, while artists, designers and computer scientists have already created a number of projects to visualize cultural media, the existing projects that I am aware of rely on existing metadata such as Flickr community-contributed tags. In other words, they use information about visual media — creation date, author name, tags, favorites, etc. — and do not analyze the media itself.

In contrast, Cultural Analytics uses image processing and computer vision techniques to automatically analyze large sets of visual cultural objects to generate numerical descriptions of their structure and content. These numerical descriptions can be then graphed and also analyzed statistically.

While digital media authoring programs such as Photoshop and After Effects incorporate certain image processing techniques such as blur, sharpen, and edge detecting filters, motion tracking, and so on, there are hundreds of other features that can be automatically extracted from still and moving images. Most importantly, while Photoshop and other media applications internally measure properties of images and video in order to change them — blurring, sharpening, changing contrast and colors, etc. — at this time they do not make available to users the results of these measurements. So while we can use Photoshop to highlight some dimensions of image structure (for instance, reducing an image to its edge), we can’t perform more systematic analysis.

To do this, we need to turn to more specialized image processing software such as open source imageJ which has been developed for live sciences applications and which we have been using and extending in our lab. MATLAB, popular software for numerical analysis, provides many image processing applications. There are also specialized software libraries of image processing functions such as openCV. A number of high-language programming languages created by artists and designers in 2000’s such as Processing and openFrameworks also provide some image processing functions. Finally, many more techniques are described in computer science publications.
While certain common techniques can be used without the knowledge of computer programming and statistics, many others require knowledge of C or Java programming. Which of the algorithms can be particularly useful for cultural analysis and visualization? Can we create (relatively) easy-to-use tools which will allow non-technical users to perform automatic analysis of visual media?

These are the questions we are currently investigating. As we are gradually discover, in spite of the fact that the fields of image processing and computer vision have existed now for approximately five decades, the analysis of cultural media often requires development of new techniques that do not yet exist.

To summarize: the key idea of Cultural Analytics is the use of computers to automatically analyze cultural artifacts in visual media extracting large numbers of features which characterize their structure and content. For example, in the case of a visual image, we can analyze its grayscale and color characteristics, orientations of lines, texture, composition, and so on. Therefore, we can also use another term to refer to our research method – Quantitative Cultural Analysis (QCA).

While we are interested in both content and structure of cultural artifacts, at present automatic analysis of structure is much further developed than the analysis of content. For example, we can ask computers to automatically measure gray tone values of each frame in a feature film, to detect shot boundaries, to analyze motion in every shot, to calculate how color palette changes throughout the film, and so on. However, if we want to annotate film’s content – writing down what kind of space we see in each shot, what kinds of interactions between characters are taking place, the topics of their conversations, etc., the automatic techniques to do this are more complex (i.e., they are not available in software such as MATLAB and ImageJ) and less reliable. For many types of content analysis, at present the best way to is annotate media manually – which is obviously quite time consuming for large data sets. In the time it will take one per-
son to produce such annotations for the content of one movie, we can use computers to automatically analyze the structure of many thousands of movies. Therefore, we started developing Cultural Analytics by developing techniques for the analysis and visualization of structures of individual cultural artifacts and large sets of such artifacts – with the idea that once we develop these techniques we will gradually move into automatic analysis of content.

**DEEP SEARCH**

In November 2008 we received a grant that gives us 300,000 hours§ of computing time on US Department of Energy supercomputers. This is enough to analyze millions of still images and video – art, design, street fashion, feature films, anime series, etc. This scale of data is matched by the size of visual displays that we are using in our work. As I already mentioned, we are located inside one of the leading IT research centers in the U.S. – California Institute for Telecommunication and Information Technology (Calit2). This allows us to take advantage of the next-generation visual technologies – such as HIperSpace, currently one of the highest resolution displays for scientific visualization and visual analytics applications in the world. (Resolution: 35,640 by 8,000 pixels. Size: 9.7m x 2.3m.)

One of the directions we are planning to pursue in the future is the development of visual systems that would allow us to follow global cultural dynamics in real-time. Imagine a real-time traffic display (a la car navigation systems) – except that the display is wall-size, the resolution is thousands of times greater, and the traffic shown is not cars on highways, but real-time cultural flows around the world. Imagine the same wall-sized display divided into multiple windows, each showing different real-time and historical data about cultural, social, and economic news and trends – thus providing a situational awareness for cultural analysts.
Imagine the same wall-sized display playing an animation of what looks like an earthquake simulation produced on a super-computer – except in this case the “earthquake” is the release of a new version of popular software, the announcement of an important architectural project, or any other important cultural event. What we are seeing are the effects of such “cultural earthquake” over time and space. Imagine a wall-sized computer graphic showing the long tail of cultural production that allows you to zoom to see each individual product together with rich data about it (à la real estate map on zillow.com) – while the graph is constantly updated in real-time by pulling data from the web. Imagine a visualization that shows how other people around the world remix new videos created in a fan community, or how a new design software gradually affects the kinds of forms being imagined today (the way Alias and Maya led to a new language in architecture). These are the kinds of tools we want to create to enable new type of cultural criticism and analysis appropriate for the era of cultural globalization and user-generated media: three hundred digital art departments in China alone; approximately 10,000 new users uploading their professional design portfolios on coroflort.com every month; billions of blogs, user-generated photographs and videos; and other cultural expressions which are similarly now created at a scale unthinkable only ten years ago.

CONCLUSION

To conclude, I would like to come back to my opening point – the rise of search as a new dominant mode for interacting with information. As I mentioned, this development is just one of many consequence of the dramatic and rapid in the scale of information and content being produced which we experienced since the middle of the 1990’s. To serve the users search results, Google, Yahoo, and other search engine analyze many dif-
Different types of data – including both metadata of particular web pages (so-called “meta elements”) and their content. (According to Google, its search engine algorithm uses more than 200 input types.) However, just as Photoshop and other commercial content creating software do not expose to users the features of images or videos they are internally measuring, Google and Yahoo do not reveal the measurements of web pages they analyze – they only serve their conclusions (which sites best fit the search string) which their propriety algorithms generate by combining these measures. In contrast, the goal of cultural Analytics is to enable what we may call “deep cultural search” – give users the open-source tools so they themselves can analyze any type of cultural content in detail and use the results of this analysis in new ways.
AN INTRODUCTION TO CULTURAL ANALYTICS
Marie-Anne Chabin is director of Archive 17 (www.archive17.fr), an archival experts and advisory firm. She first gained experience as an archivist in the public sector (Direction des archives de France) before deciding to broaden her professional experience by joining a private firm as an EDM consultant, and later the Institut national de l’audiovisuel (INA).

In 2000, she decided to set up her own firm, Archive 17, as an expert in records management and electronic archiving, to help companies to promote records management policies and guidelines, retention schedules, requirements for records management systems and training. In 2007 she developed the concept of “retention categories” and a new method, named ARCATEG, used to manage records in a very simple and efficient way.

Marie-Anne has been recently appointed as an associated professor at the Conservatoire national des Arts et Métiers for the records management training program.
I AM VERY pleased to be involved in this exciting debate about preservation and selection of audiovisual archives here in Stockholm, first because it reminds me the quite interesting two years I spent at the French Institut National de l’Audiovisuel (INA) ten years ago, second because I guess, as a specialist today in records management matters, that the need of selection is more or less the same for all kinds of documents (with different strategies) as an answer to the huge growth of information in the digital world.

DIGITAL AUDIOVISUAL ARCHIVES

Before going further regarding selection, it should be useful to comment a little these three words.

“Digital” refers to the electronic format and medium and the associated tools.
The word make you know that what we are talking about is made of 0’s and 1’s and not of former technologies. But everything is becoming digital so that the word will not be necessary in a near future.

“Audiovisual” means sound and/or image, in opposition to text. The adjective was first limited to radio and television, public and private national channels, then the local channels and today the world wide web. The most remarkable fact of the history of audiovisual during the last decade is the wonderful success of social networks. I am not a member nor a regular user of Facebook or YouTube but I must confess that I am fascinated by the development of such communication means, mixing text and audiovisual. This success suggests that audiovisual could become a common language, a kind of Esperanto of the 21st century, allowing people to exchange information and communicating better than words of a single language could do. Facts and comments on the main events of our world, related to environment or political matters, are shared through images and videos, posted on social networks both from official and non official sources. These videos may be of bad quality, it does not matter. The content, that is what you say, what you show, is often more important than the technical work. There is no art; it is just a way to speak to a large audience, to an unknown audience. It was the same in the ancient Roma when somebody decided to explain something in the forum to the people around and people were free to listen or to answer, or not.

Due to the same technological progress and associated practices, the part of audiovisual is also growing up in the organisations as records of evidential value, sounds and images replacing text for security controls or in knowledge management processes.

“Archives” has several meanings. Traditionally, archives are those records kept permanently after retention period is over, for historical and cultural reasons. The main part of audiovisual archives, as television production, are not records of evidential value likely to be used in administrative or commercial litigation and they are considered as “archives” just
after broadcasting, but the aim of preservation is the same, and includes also business purpose. The word Archives has an additional meaning in the digital world, i.e. any library or collection of resources in electronic form, as in the expressions “Open Archive Initiative” or “Open Archival Information System”.

A characteristic of archives, whatever the meaning is, is that some are “born” archives while others get their archival status after a while thanks to a selection process, so that they become archives “by baptism”. Whatever the media is, some documents are to be considered as archival material as the very moment of their creation; for other documents, a period of time is needed to validate their archival value. Archives by birth are those documents which content or provenance avoid any question about their retention neither from their creator nor from the institution in charge of preservation. Archives by baptism are those materials which get the chance to be retained as valuable material for the future thanks to archivist’s knowledge and personality, or simply to human and technical resources and opportunities.

**WHAT DOES SELECTION MEAN?**

Currently speaking, the selection refers to the choice of some elements among an existing population (seeds, pupils, sportsmen, etc.) with the purpose of improving something such as quality or performance…. The aim must be defined clearly before selection is performed. Then, those who can be involved are identified, compared and taken in or out. Those who are not selected go on living normally or try to be selected in another competition.

In archives management, selection may have the same meaning, i.e. identify the best material for an exhibition, a festival, a project and so on. The main purpose is not to improve a process (it could be a cultural process
in that case) but to give the best photograph of the past.

There is another way to define selection, focusing the global issue, the management of a complete collection, and not only the result of the process or the use of this result. Facing an existing amount of archives, the process of selection means to choose a part of them, according to pre-defined criteria, and destroy the other part. The difference is that in the first case, some content are retained for a specific purpose, and in the second case, some content are non selected for any use and there will be no other chance to use them. Here, the opposite of selection is destruction.

We could compare the work of archival selection to the work of a sculptor, who can use terracotta to build a figure by adding small parts of clay together, or who can use a stone and creates a form by taking away useless parts of it.

Generally speaking, selection is either human or natural. Human selection is performed by the archives managers, on the basis of professional criteria. Natural selection interferes as the result of time and history, responsible for lost, destruction, leaving out, neglecting...

Audiovisual and digital world do not really change the principle of archival selection: audiovisual technologies create new contents generating new interests according to a new public; digital information changes the possibilities and constraints of preservation and access, but the principles remain. Regarding preservation, the difference is that, at least today, you cannot leave audiovisual tapes on a shelf and came back fifty years later, as we did with parchments and paper to evaluate if there are worth to be kept longer. The tapes will not last so long and I am afraid that technological progress will never save the orphan material.

WHERE DOES THE NEED OF SELECTION COME FROM?

The need of selection (selection to keep or selection to destroy) arises from
different positive and negative contexts, but there are more or less four main trigger events.

The first one is the cost of preservation and access. The cost of preservation of digital audiovisual material depends on both on the volume of material (which is increasing continuously), and on the cost of technology (which is decreasing). The characteristics of formats and medium used at the time of creation and at the time of archiving are important as far as the cost of migration, even decreasing, is still high. Then who pays? The owner, the community, the users, everybody? But what should each of them pay for? It would be very interesting to link the cost of collecting archives and the cost of preservation with the knowledge of available archives and the real use of them.

The second reason for selection is the quality of the material. There are different levels of quality, both from a content point of view and from an aesthetic or technical point of view. It is obvious that some archives are of less interest than others and it is obvious too that the most valuable archives should be kept first. But the level of interest is more an opinion, depending on culture and generation, than an objective fact. Do we want to document the most seen sequences or, on the contrary, those that the public did not understand, out of the fashion at a time? Do we want to preserve only the material of good audiovisual quality or also the unsuccessful contents and make what we could call an “anti-selection”. It is reasonable not to keep too much copies of the same content, but somebody may be interested in retaining precisely a series of similar content, just to analyze the 1 percent difference. And so on. Anyway, whatever you keep, you have to document the material and the context of creation/broadcasting. At this point, the lack of documentation is an objective criterion to leave a material out of the selection.

The owner policy is also a criterion, in connection with the others. Unfortunately, the owners of archival funds or the organisations who have
addressed the archival issue and defined a policy, at the level of the archi-val material or at the higher level of information management, are not many. Most of time, archives are accumulated and managed in a very practical way, on the basis of logistic problems and users needs. But it happens more and more that, facing the amount of material and the associated costs, organisations try to sketch out archival policies including the principle of selection.

One of my customers, member of the Executive committee of an international luxury group, told me a few weeks ago: “We need to keep essential content, the best part of the production. We never reached comprehensive collections. And, even if we had the means to keep everything, what for should we keep it? We want to present a view, a feeling of our activities. It is more significant to keep five looks which have been chosen and documented than to keep ten looks without any choice.”

The last trigger event for selection in this list is the right of privacy, at the time of broadcasting, and the right to oblivion afterwards. Facing the possibilities of technology, a lot of people feel free to create and broadcast private content (TV-reality, video disseminated through social networks) and are not aware of the right to re-use it or not after some months or years. There are some international or local laws but the phenomena it still recent and the regulation is not yet complete. A related issue is the necessity of archival principles in managing audiovisual material in order to avoid the manipulation of images and video sequences out of their creation context. The “non selection” of some too private contents should be considered as the most respectful approach in some cases.

**WHEN DOES SELECTION TAKE PLACE?**

The concept of human archival selection seems to be relevant after the
creation of the audiovisual documents (that is when technical problems arise), at the time of creation/broadcasting/dissemination or even before creation/production.

BEFORE THE CREATION / PRODUCTION

The idea of selection before creation could sound strange, unless you consider that selection means also deciding which content must be created as relevant archives, according to the need or will of memory or evidence, particularly when there is a risk that this content does not exist or will not be retained elsewhere. This is also valuable for records: the creation of some audiovisual records must be linked to the will of completing the collective memory as well as the need to be compliant with the regulatory environment or to manage risks. In this case, the decision of selecting a content to be created just for archives, without any other immediate purpose, must be made or agreed by the creator himself.

A good example of this “pre-selection” is the film the former French president, Valéry Giscard d’Estaing, decided to make before his election in 1974. As a candidate – and nobody could be sure at that time that he will be the winner – he asked the French photographer and film maker Raymond Depardon to stay with him a couple of weeks during his campaign before the vote and to make a film. The film, named “Une partie de campagne” (there was already a Jean Renoir movie of that title), shows the aristocratic candidate Giscard d’Estaing visiting modest families in small towns or talking with his advisors when listening to the speeches of his challenger François Mitterrand. The film remained secret during 25 years before the no more president decided to make it public ten years ago. To the question of a journalist about the reason why he made this film and why he kept it secret for so long time, Valéry Giscard d’Estaing gave this interesting answer: “I did it for me and for the archives”. There is in this
sentence a part of selfishness but we have in the archives, from the oldest periods of History, such historiographical material which are of great interest.

AT THE TIME OF CREATION / BROADCASTING / DISSEMINATION

Selection at the time of broadcasting is the most obvious way of selection for audiovisual material, at least for TV programs. It refers to the combination of two criteria: the content, including topic, redundancy, rights, and the audience, which a precious metadata element when analysing or reusing the subject.

Retention of the main programs, that is the most often seen by the public; or the programs the sponsor has chosen as the best ones, or those that professionals have retained for technical, legal, business or cultural reasons.

An example of this kind of selection is the choice to retain, for “world memory”, the ten most viewed videos on YouTube at a time, among them, this month, Lo que tú Quieras Oír (‘What you want to hear’), a short film starring Fatima Baeza, in Spanish with more than 7 millions views. If the statistics on social networks are reliable (to be checked), it is really worthwhile to keep in the archives what so many unknown people have shared through the Internet. The remaining question is: which organisation would be in charge of such a worldwide content?

AFTER THE CREATION

A couple of years after creation, the evaluation will be different, thanks to archival distance and experience; in addition, time allowed to take into account external criteria such as new needs among users and customers.
From this point of view, it is obvious that a video of the beaches of Phuket (Thailand) made by a tourist in 1984 and posted on wat.tv has an enhanced value for selection after the 2005 tsunami which destroyed the landscape.

**HOW TO PERFORM SELECTION DECISIONS?**

Generally speaking, there are two ways to perform selection process: automatically or specifically. If the characteristics of the material are identified at the time of creation,

If the content and the context are too specific, a human intervention is needed to appraise the situation and validate the selection.

Automatic selection can be processed but adequate tools or by a dedicated team, on the basis of mechanical criteria, with the purpose of keeping the content of huge and redundant series through a sampling of the material:

- One out of 10, 50, 1000
- the first and the last of a series
- one per month
- the most viewed
- the creators included in a predefined list
- the content including such image or such word

If the material cannot be sorted out automatically on the basis of practical elements, the involvement of somebody is required to make the selection at the time of capturing or archiving the material. The archivist or the expert (or the group) in charge of selection will identify the events or topics to be documented for collective memory and look for the most significant
documents according to users’ needs, future use, as far as you can have an idea of it…

Both automatic selection and appraisal of content are likely to be reviewed after a while, a kind of “second selection”.

In any case, the work will be documented and the selection should be managed through indicators and scorecards to document and follow up the selection process:

- are the requirements of the selection policy fulfilled? if not, why? difficulties linked to identification, to capture, to cost of preservation? how to improve selection policy if there is one and, if not, to produce an appropriate policy or at least a guideline

- is the access to those archives significant? what can be done to improve the use of this cultural heritage?

- are some collected archives to be deleted if assessment shows it doesn’t fit with requirements?

**CAN WE SKETCH OUT SOME STRATEGY MODELS?**

Every organisation is free to manage its archives with its own rules and we can imagine a lot of different manners to implement selection. Among all the possible strategies, I think there are four main models, according to the main environments where selection is likely to take place.
1. ORGANISATION APPROACH

If the archives are managed within the public or private organisation as a corporate memory, I guess that my 75-25 theory could be relevant. My experience has suggested to me a kind of archival law according to which, regardless the activity or the medium, 75 percent of corporate memory come from records and 25 percent from personal or external sources; these 25 percent of miscellaneous archives should be collected from regular sources or picked up here and there, depending on events and people concerned. A specific classification should be created for those 25 percent as there is classification scheme for records with disposition requirements identifying records of cultural value. Selection in this case means to define what are best document types and quantities to complete the organisation memory, in addition of records.

2. BUSINESS APPROACH

When the archives are managed by an archival institution or dedicated company, particularly in the case of audiovisual collections, archives may be considered as a commercial product and the selection of archives applied according to this context: what kind of material has a business value at the time of production and what kind of material has not but could acquire such a value in the next or further future? Business requires both a good quality of material and identification of potential customers, for example those people shown on the video and their family in relation to a public event or popular program.

In this scenario, the return on investment is the key factor and there are two strategies: on one hand, a recent content can to be spread to a large public in the next future; on the other hand, some archives will gain value over years and become a rare and expensive product.
3. INDIVIDUAL APPROACH

More and more collections are organised with an individual or personal approach. Archives are collected or created to enhance personal memory and family history when keeping private videos or public videos showing private life to be shared with the other members of the family or friends. In this case, the quality of material is not an important criterion; the most important is the identification of the creator and the family. The interest of these archives is likely to decrease after two or three generations, when people directly involved are dead; unless the content refers also to common events, such as postcards sent during the First World War for the last decades, but that refers to the next category.

4. COMMUNITY APPROACH

The fourth model, which is rather new because linked to the use of the Internet, refers to the replacement of territories, which are the normal scope of an archival collection (national, local), by communities. The French philosopher Jacques Derrida has written some pages of interest about the growing gap between territories and communities: the worldwide network doesn’t care about territories, states and institutions. On the contrary, it enhances specialized communities based on common ideas, religion, ways of life, hopes or hobbies of their members, regardless of territorial provenance.

These communities have generally a voluntary attitude for building a new identity outside territories and institutions. The selection here requires the identification of the most representative production within this community to be documented, combining selection of sources and of relevant material to be preserved for identity and history.
CHANGES TO COME IN THE FUTURE

Internet and the social networks are the main change of the last decade and it is difficult to imagine other changes of the same size for a while but all these trends in the field of audiovisual archives will probably go on within the next five or ten years:

- the part of audiovisual in information world will grow up, whatever the type of activity and management concerned: records management, content management, portals, social networks…;

- awareness of good practices will be larger in public and private organisations: the main issues related to the preservation of digital material will be shared by all the stakeholders, in public and in private sectors; redundancy will be managed on the basis of scientific methodology; the need to destroy audiovisual content for privacy reasons will be understood and respected;

- multiplicity of individual initiatives to keep public content for a while, for personal use or deliberately to make them become better over years, like wine in the cellar… In a way, technologies and the place given to entertainment in our societies allow everybody to become an archivist and to collect information;

- technological progress is going to offer new possibilities, new means to help not only preservation but also appraisal and selection through bulk processing. Search engines based on “non text” elements (i.e. forms, lines, colours, contrasts) will be used to reach and compare contents that will remain unavailable and unmanageable without the support of appropriate tools. This is particularly true for those who are still curious and cannot be satisfied with the only
information they receive without searching anything. The benefits of this “technological serendipity” will certainly be enhanced in the near future, particularly concerning selection of audiovisual material on the web.

Beside changes, it is also important to identify some permanent components in archives management and the attitude of the users.

It is noticeable, for example, that historians analyzing archives accumulated through decades and centuries often review the same material with new questions according to their generation, and the result is normally so relevant that archivists wonder sometimes why they keep so huge quantities of archives if only a small part is searched for. As far as I know, only ten or twenty percent of the documents kept in the French National Archives are really searched for and used: it is so for medieval archives, it is true also for the 20th century. Why should it be different with digital medium?

Of course, in the paper world, the lack of finding aids and the difficulty to retrieve scattered information may be an explanation of these ten percent and somebody could argue that the digital technology precisely allows improving the availability of archives. The long tail theory let us think for a while that it will be different with the Internet but at the end it is not. My feeling, a philosophical feeling if I may say, is that the 10 percent phenomena will last, because it is not a matter of technology, but a matter of human memory and of relevancy of the sources. The historian Jean-Claude Robert, of Quebec, once asked the question: “Do we really have a best knowledge of the past when using a larger quantity of archives?” and he demonstrate that the answer is “no”.

Another fact is that, whatever you keep, there is always somebody to ask for something else, something of the poorest interest, or for something that has never been created, sometimes only because looking for missing information is in a way more exciting.
CONCLUSION

It appears that, whatever the technologies and the volumes of information are, archival selection is unavoidable, sooner or later, to build proper archival collections.

Selection is unavoidable for professional, political and technological reasons and it could be a sufficient reason to accept it. But I think that if we had the possibility to keep everything, it would lead to suffocation: could we survive without forgotten anything? Selection and oblivion are a matter of social breath. The risk is that, deciding to keep this and to reject that, some further event will make us regret our choice. Well, could we live normally without regretting anything? As explained Joachim du Bellay, a French poet of the 16th century, regrets are a part of human being.
Dr. Robbie De Sutter received his Ph.D. degree in computer science from Ghent University, Belgium, in 2006. His research topics included video coding technologies, usage context modelling and negotiation, and content adaptation. As a researcher at VRT-medialab, he is managing the information management research group. This group researches and develops, in collaboration with universities and other companies, technologies and architectures to create, adapt, integrate, and exploit audiovisual descriptive metadata.
“CONTENT IS KING, Metadata is Queen”. The audiovisual entertainment sector adheres to this old adagio, but it is a fact that without (correct) metadata content is unusable. Traditionally, the creation of metadata is a task of archivists: they are responsible for creating high quality metadata, including technical, descriptive, administrative, and even legal information.

Due to the fact that an archivist is constrained in manpower and time, the created metadata is limited. Retrieving audiovisual material is therefore hampered, forcing a program researcher to spend a lot of time to find the desired material for reuse. For example, if no time-based metadata (which associates information to a particular time range in the audiovisual asset) is available, a program researcher is forced to browse through the entire video rather than being able to jump directly to the desired scene.

To address these issues, a new research project – “Memento” – is under-
way. Memento’s goal is to increase the quantity and quality of metadata by developing tools, systems, and guidelines that make finding and reusing audiovisual material more efficient.

**APPROACHING METADATA**

Memento will focus on the augmentation of the quantity and quality of metadata by gathering and exploiting the information created in the production environment of a media enterprise and by employing (Internet) user communities. Next, Memento will develop new applications that will make searching in audiovisual collections, thanks to the added metadata, faster and easier.

Within the production environment, Memento will provide solutions to collect more information about the produced material by optimizing the information flow during the production process. The project will investigate how the existing production information (such as, scripts, scenarios, subtitles, rundowns, editing decision lists, etc.) can be harvested and linked to the finalized media material. Most of these production information sources are currently unstructured (e.g., plain text documents, Excel or Word documents, etc.), archived but unlinked (i.e., no reference to the resulting or original material), or not available in electronic form (e.g., handwritten pieces of paper). Making the information sources available in a structured way is a prerequisite for more efficient search applications. For example, rundown information created before and updated during a news broadcast provides valuable time-based information (such as teleprompter text and text captions), however this information is currently not linked to the media material and is not retained after the news broadcast. Memento will first create an overview of the information sources available during production and will investigate their potential added value if the information they contain would be available in structured form. Next, new proof-
of-concept applications will be developed for those sources with high added-value. These applications will be optimized to support the end-users in the tasks they want to accomplish, and at the same time they will capture the information in a structured, exchangeable, and interpretable format.

THE SECOND APPROACH

Memento’s second approach to gain more qualitative metadata is by employing tools which extract information directly from the audio and/or visual content stream, such as: shot cut detection, face detection and recognition, object class recognition (vehicles, pedestrians, etc.), scene recognition (indoor, city, landscape, etc.), and so on. Although feature extraction tools are not new, most of these tools are still in their infancy and/or exclusively available within the academic research community. Memento will optimize some of the available feature extraction tools for usage within a broadcaster environment. For example, current face recognition tools can distinguish between a few dozen faces and this only after intensive training. However, in order to be useful within a media enterprise, they should be able to identify several thousands of different faces. Furthermore, since new faces are added daily, their training should be very quick and easy and preferably for the most part automated. In addition, feature extractions tools should no longer work in isolation, but instead take the available context into account as well. For example, if script information is available, the number of people on stage is known and limited; as such the use of this information should improve the precision with which the object recognition tools correctly recognize an actor. Finally, these tools must be “media-aware” in the sense that they return time-based information (e.g., identified person X appears from time-position A to time-position B).
THE THIRD APPROACH

The final approach to increase the quantity of metadata and their quality is to explore how an (Internet) user community can be involved to create new information about the audiovisual material. The core idea of Memento is to make content available to a user community by using existing Internet (cloud) services or enhanced IPTV technology so the community can freely dispose of and (re)use it. Next, new hybrid (television and internet) formats will be invented and brought to realization. These new formats will allow us (a) to reach the chosen community and provide them with material, (b) to stimulate them to add descriptions, tags, and comments to the material, and (c) to harvest the implicit and explicit time-based community generated metadata. Experience and guidelines will be built up on how to reach and deal with communities and how to motivate and stimulate them to add information about media material.

Within this third approach, this “user generated metadata” will be collected in different ways.

Implicit feedback can be collected in combination with interactive content. Here, the user can interact with the content in a natural and typically leaned-back way. User actions, such as zapping away, skipping a certain fragment or recommending a certain fragment to a community, can be used to harvest important information about the fragment. In the analysis, several elements are crucial to make meaningful conclusions. Information about the user and his preferences is almost inevitable (what are the topics in which this user is interested). In case of community-interactions, also knowledge about the field(s) of interest of this community is required. In this approach, feedback information from as many people as possible is required to assure a sufficient quality of the harvested metadata. In the media delivery area, research concerning the scalable delivery of interactive and customized content is required and as well as further research in areas such as interactivity and personalization of content, the ability to support
new TV-formats and the collection and processing of user actions. Based on fine-grained metadata that is time-based, new TV-formats that were previously not feasible due to the lack of metadata and the rigid structure of production process can now turn into reality. In this way, metadata also perform the role of an enabler to deliver more interactive and personalized content towards the end-user.

Using explicit user feedback, the user will actively participate and will provide feedback either on a personal basis or embedded in a community. Possible ways to provide feedback could be tagging, commenting or even freely annotating available content. Employment of user communities typically results here in metadata on a semantically higher level than feature extraction tools and solves the archivists’ lack of time and manpower issue, but implies the problem of quality assurance.

Before the information (originating from the optimized media production tools, the feature extractions tools, and the user community generated metadata) can be used within the business environment, it must be processed. In particular, undesirable information must be filtered (e.g., spam tags), multilingual information must be handled as well, and the metadata must be transformed to the specific ontologies or thesauri used within the media enterprise. Furthermore, storing and indexing the expected huge amount of information requires new techniques seeing that the classical “row-based” databases are no longer suitable.

The processed information will complement the information created by the archivists (if any) and provides additional point-of-views on the same content. The added-value of the increased quantity of metadata (and the usefulness of the different methods) will be demonstrated in Memento by efficient search applications optimized to search within an audiovisual catalogue. These applications will be suited to the needs of the program researchers and will exploit the additional collected information from the various sources and approaches. Contrary to classical text document indexing and search, the audiovisual search engine in Memento will be
“media-aware”, hence it will be able to deal with time-based information. Advanced search queries including the time-based information – such as finding a clip where person X appears before event Y occurs – will be developed. Because of this, program researchers will be able to express advanced and goal-oriented search queries to find and retrieve the desired material.

**IN CONCLUSION**

Memento provides solutions (applications, tools, techniques, and guidelines) to enable users (professional media enterprise program researchers and archivists, semi-professionals, and home users) to find and retrieve audiovisual material more quickly and accurately thanks to a larger number of high quality information about the content and optimized search applications. This results in more refined results in a shorter timeframe and improves the overall efficiency and productivity of program researchers.

The project “Memento” is submitted as an IBBT/IWT ICON (www.ibbt.be) project and is currently under review. The project consortium consists of several academic and industrial partners, each selected for their particular competence and complementary. If the project is approved, the activities will start on January 1st, 2010 for a period of two years. The outcome of the project will be available in the form of several publicly available demonstrators and (scientific) papers.
Since 2003, Anna has been an Archive editor at the TV Archive at SVT. From the beginning of 2008 she has been working in the SIRI Project, educating and advising the Program Production staff in how to add metadata to their digital program files. Anna has a master degree in Library and Information Science from Borås and has written a Master’s thesis about standardising metadata in public service. Prior to that she has been working as a journalist and archive editor in different media companies.
5 METADATA QUALITY CRITERIA

IN THIS PAPER I will present a project aimed to develop quality assurance of the production and transmission metadata at SVT. The project is closely connected to the digital production workflow of SVT and the new ways of archiving in the company. Since the beginning of last year I have been working at the department of Rights Coordination at SVT, together with our chairman of the Media Management Commission, Eva-Lis Green. My main task has been to teach the program makers to add their own metadata into our program and archive system.

BACKGROUND – CHANGING SCENERIES...

First I will give you a background of the changing sceneries at SVT. In 2001 the news and sports production at SVT became tapeless, and in the
fall of 2007 the rest of our TV production followed, which also meant a changing workflow which three main issues:

1. All metadata should be collected during production

2. Program makers should be responsible for complete delivery of program files and delivery specification (including version data, technical data, rights, content, basic data and music)

3. After program delivery there should be a “new owner”

... CHANGING ROLES

Shortly after this change of scenery, in the fall of 2007, one of my colleagues, Berit Greider, flew to the north of Sweden to take part in a pilot project, of course led by Eva-Lís Green, where her main focus was to teach three program teams to add program metadata into the digital production system. All the programs are produced digitally, but in different ways. The first program is a one hour lifestyle program, with studio interviews and reportages about health, culture, cooking, travels and so forth. It has a running order where all the files of the reportages are posted and easily available for the program makers. The second is a nature program, with the host introducing different reportages in the environment. The third program deals with consumption issues, and includes studio interviews interluded with reportages.

Berit started off with an introduction to all the members of the teams, explaining that this change was decided by the board and compulsory. After that she met all the members of the program teams and had “hands on” lessons with each one of them at their computers for approx. one and a half hour to start them off. The main focus was to find the right time
to add the metadata in the workflow of the different programs, and also have it done by the person with knowledge of the material. Depending on the program it could be the reporter, the producer, the “redigerare” or “scripta”. The lessons included a program description, key words, reuse rights information and “arkivbeslut”. The program teams all found their own way to add metadata in their own workflows. A few months later, after an evaluation of the pilot project, the board decided that this was a possible way to continue.

**METADATA DELIVERY: SIRI PROJECT**

And that was the start of the SIRI project where I am working now together with Berit and three more colleagues from the Archive. Eva-Lis is leading the project, which name means “add the right information”. Our main focus is to teach the program makers how to add metadata to their program files, but we also give support and feedback a long period afterwards. Our project does not include bulletins, daily news and sports programs, but everything else from drama to documentaries. Today our project is well known throughout the organization, and although there are still a lot of problems to solve, lack of time being a main issue, we have managed to make the process of adding metadata part of the production workflow in most of the program projects.

In the beginning of June we also took a leap towards what we call “Complete delivery” of our programs. It means that the program makers have to deliver their metadata at the latest one week after transmission to a Program Delivery Center. If they don’t deliver in time, they will get a reminder from the center. A week after that, me and my colleagues, now in the function of a Central Program Reception, will remind them again together with the program director. If the metadata is not delivered after one month, there will be charges against the projects.
NEW NEEDS AND CHALLENGES FOR METADATA QUALITY CONTROL

Today there are hundreds of program makers adding metadata to their own program files in our production and archiving system called META. They have all been given user manuals, rules and regulations, but of course the amount and quality of metadata they put in varies a lot, also depending on the type of program they produce. Needless to say, there is a new urge for quality assurance in this kind of workflow. But we also tell them that in a digital environment they don’t have to describe their programs the way an archivist did before, because program files are much more available in a digital environment. Instead we are emphasizing that they have to put their effort more on detailed information of the re-use rights. A short description of the program together with keywords and the names of all participants are sometimes enough.

The SIRI project made it necessary to be more and more exact about what kind of metadata is needed. Not only what we used to do from our old cataloguing rules, but to really think what is needed in the new environment. We have to define more exactly: What have to be captured and kept? How do we define the lowest limit for different programs? What is good enough metadata quality? We also need to find a way to control and quality assure our metadata referring to our collections. We figured out that we needed a scheme or some structured way to go about since we no longer can rely on the archivist checkpoints.

QUALITY MEASURES

Since I had recently written a Masters thesis about a metadata project at SVT I could contribute with some measurements of metadata quality, which I had found during my research. Our main source of inspiration has been Diane Hillmann at the Digital Library Research Group at Cornell University, USA. In 2004 she published a very interesting anthology
together with Elaine Westbrooks called “Metadata in practise” edited by the American Library Association. There she wrote, together with Thom-as Bruce, one chapter called “The continuum of metadata quality” where they defined general characteristics of metadata quality on an abstract level, which partly is built on a framework developed by Statistics Canada. When I showed these quality measures to my project manager Eva-Lis, she found them helpful even for our purposes.

Diane Hillmann and Thomas Bruce examined seven of the most commonly recognized characteristics of quality metadata and made consider-ations of their value in an abstract way. At SVT we have remodeled their work for our own collection-specific purposes. First I will describe the seven quality measures that we are using:

1. Completeness – metadata should describe the programs as complete as our economy allows.

2. Accuracy – metadata describing the programs should be accurate, correct and factual.

3. Provenance – emphasizes the importance of knowing if and when metadata has been changed, updated or added.

4. Conformance to user expectations – emphasizes the importance to communicate the “choices of metadata” that has been made and help the users by providing user manuals and spelling lists.

5. Logical consistency and coherence – metadata in different systems should be added consistent and presented in a consequent way.

6. Timeliness – metadata should be added as soon as possible after the program file is available in the archive system, to avoid empty files.
7. Accessibility — all user groups should have access to the archive system without extra costs and with standard desktop tools.

With the help of these measures we have formulated criteria and answers about how to improve our metadata and have also suggested activities (meaning what has to be done) and indicators (meaning how and where this should be checked) for each criterion. I will give you some examples of our work:

COMPLETENESS

The first example concerns completeness, and the criteria is: Does the metadata elements completely describe the programs? Our answer is yes, for broadcasted programs. (We haven’t started to investigate the web yet.) The activities that we are planning are continual updating of user manuals and also continual check-up of metadata schemes. The indicators that we have suggested are information system application profiles and documentation of decided standards.

My second example is also in the area of Completeness. The criterion is: Are all relevant kinds of metadata used to describe each program? We think it is, but we have to develop spot-checks etc.

We want to create more standard outputs for check-up both for production and broadcast metadata and define which metadata should be compulsory, for example rights metadata. We suggest that the indicators, meaning how and where it should be done, should be field statistics and continual check-ups. We also need an approved delivery to the Central Program Reception, which is a new function for metadata and program delivery at SVT which has just started.
ACCURACY

The next two examples concern accuracy. The first criterion is: What has been done to secure a "good enough" level of metadata? The situation today is that we have educated all program makers, excluding the bulletins, daily news and sports programs. We have also defined in user manuals which metadata should be kept and suggested levels for different kinds of programs. We think that what we have to do is define minimum and maximum levels for different categories of programs and also to develop control functions like spot- and spell-checks.

The next example is interesting when you have to check a large amount of metadata. The criterion is: Are empty metadata fields accepted, and have they been used correctly? Today the system accepts empty metadata elements, and very few of them are compulsory, for instance re-use right is not compulsory, which is a problem since all program files should include information about restrictions to re-use the material. To change this we have to define more compulsory elements (fields) and develop reports from information systems. The indicators we suggest for this example is search tests and field statistics.

PROVENANCE

In this example we explore the provenance, eg. the history of the metadata. The criterion is: Who is responsible for creating, extracting, or transforming the metadata? The (program?) project manager is responsible for the metadata that his/her team add into the system. When they have delivered their metadata after the first transmission, and it has been accepted by the Delivery Center and the Program Reception, the Rights Department will be responsible. To improve this we have to define the roles of the Rights Department and the Program Reception of Complete Delivery.
A new definition of the role of the project manager has been defined at SVT during this spring and will be implemented during autumn 2009, and is also emphasizing the enhanced role of the project leader to take responsibility of all the administrative parts of his/her project.

I will give one more example of provenance. The criterion is: Who has created and extracted the metadata? In this case reliable log files are available in the transmission system, but not in the production and archive system. We have to develop reliable log files with user-id’s even in the production system.

CONFORMANCE TO USER EXPECTATIONS

The next quality measure to explore is conformance to user expectations. The criterion is: Does metadata describe what it claims to? Our answer is yes, in principle, but it is not delivered in a way that is convenient for the user. Metadata may also be missing if the program file is corrupt or has been deleted. We have to develop a user-friendly search interface and the search engine has to be able to exclude and include categories of metadata that is convenient for research, like subtitles etcetera.

I will give one more example of conformance to user expectations. The criterion is: Are compromises documented and in line with organizational expectations? There are many compromises in the systems as well as format changes (Beta, HD tape and files) that is not completely documented. Therefore we have to improve documentation both within the SIRI project and in the administration of the production and archiving system (META). It is also important to communicate the problems we find in the production teams to the system administrators. Indicators for this example may also be user studies.
LOGICAL CONSISTENCY AND COHERENCE

For the next quality measure we have suggested one criterion: Is the metadata consistent throughout the organization? We can say that continual work is going on to maintain the standard of all program information. To succeed we think there is a need to develop a common standard for all the information exchange in our company. Indicators will be our documented metadata standards for different systems.

TIMELINESS

Next example concerns timeliness and one criterion is: Has metadata been delivered on time? We haven’t been able to answer this question properly, since the organization for complete delivery has just started. We have the regulation, but we have to wait until the end of this year before we can evaluate the outcome.

In another example of timeliness we have formulated this criterion: Is metadata updated regularly when resources are changed? The answer is yes, for instance repeats are managed in a separate function. There is also a special management and regulations concerning migration to digital files. Updating of rules and regulations are still necessary to maintain a good metadata quality. The indicators we have suggested is also continual check-ups.

ACCESSIBILITY

The last examples concerns accessibility, and the first criterion is: Is an appropriate amount of metadata fields for users of the organization being used? We think that all necessary metadata is available, but it is not pre-
sented in the right way to make the research trustworthy. For instance it is not possible yet to search by transmission date in the production system, which is highly wanted from different user groups.

Another criterion from the same area is: Is the metadata affordable to use and maintain? Every program project manager has the responsibility to maintain a good metadata standard in the systems. We have noticed that the program teams often complains about lack of time, and at the end of different projects there has often been a lack of metadata. It is important that the SIRI project continues to follow up and evaluate these problems. A possible indicator could be a consequence analysis.

**WORK AHEAD**

This is where we stand today. Our next step will be to communicate to the management the need for metadata quality checking and suggest a way to do it. Hopefully this will lead to some decisions. We also need more requirements for system reports. (Today we just have a few.) And we also think we have to cooperate more with the information system administrators to improve the systems and make them more user-friendly.

Of course we don’t have any answers yet, as all these changes are depending on the success of the SIRI project. Before we can rely on spot-checks for instance, the production teams must deliver their metadata to the Delivery Center and Central Program Reception in time and with good enough quality, and we still have a long way to go. But hopefully our quality criteria will be an aid along the way.
Vicky has worked for the BBC for 12 years, currently holding the position of Archive Manager, Digital Library BBC Scotland. Vicky was part of a team that developed the first digital library project within the BBC, working with production teams and developers to establish a fully integrated file based production, transmission and archive workflow. She spends her time developing digital archive policies and managing online storage and preserving the tape based archive.
BBC SCOTLAND’S OUTPUT covers News & Current Affairs, Sport, Drama, Factual, Music & Arts, Entertainment, Children’s & Education programmes. We produce on average 1300 hours of new television programmes and over 10000 hours of radio programmes a year and we make content for local Scottish services as well as delivering content to the network channels for broadcast to the whole of the UK. **BBC Scotland follows two main archive policies, the BBC Corporate Management Policy and BBC Scotland’s Online Archive Policy.** The **BBC Corporate policy has a four fold process** – a high level retention policy based on set criteria.

- For research, such as content that cannot be sourced elsewhere in books, periodicals or the internet, capturing specialist contributor information, to show presentation styles, to show how technology has changed.
• For re-use or commercial exploitations, such as repeat TX, stock-shots of places that could be re-used, interviews with politicians and celebrities, also for evidence and educational re-use.

• For legal purposes, Broadcasting/Communications Act, Freedom Of Information Act, Data Protection Act, Litigation and Copyright and ownership (to ensure the correct disposal of content)

• Heritage/Historical – record of the BBC’s output, a record of society (demographics, geography, social issues, language, fashion, attitudes, morals, lifestyles, values and culture) and as a record of world events and a history of broadcasting.

The policy has a high level requirement to keep all transmitted content for five years with some genres of programming being kept in perpetuity and others selected and disposed of after the 5 years period. The focus of the policy is Tape based, used for transmitted material with less emphasis on un-transmitted material.

The BBC Scotland Online Archive Policy, defines in a new digital environment how content is managed in a dual running digital and linear environment, with priority given to content for re-use and transmission. The focus is on online and highly re-usable content with less emphasis on transmitted material. Decisions are made based on the core activities of production departments and what should be most accessible digitally on BBC Scotland desktops. We still have to meet the corporate objectives to retain material for five years post transmission but they can be fulfilled by keeping material on tape rather than digitally, or by having shorter retention online but a longer retention on tape. This policy will be reviewed regularly in line with the user and business requirements.
BBC SCOTLAND: DIGITAL ARCHIVING

BBC Scotland has fully integrated end to end digital production since 2007. We are pioneers in the BBC with a fully functional digital archive system on 750 desktops with tiered storage attached using online high resolution storage and data tapes. Legacy archive is being digitised on demand and a process of dual running across linear and digital is in operation.

The digital archive is a powerful resource in BBC Scotland’s new toolkit. It allows users to search for, view and listen to content at their desktop, select content to be reused and request content to be transferred to editing tools; Avid or VCS. The current physical archive can also be uploaded on request.

Designed by Ardendo/IBM the Digital Library:

- Maximises use of the BBC archive
- Saves research and cataloguing time
- Avoids duplicated originations
- Minimises tape storage space
- Removes tape transfer time
- Allows better rights tracking
- Minimises the need for re-shooting material
- Safeguards the integrity of the archive by storing material in non-proprietary formats with mirrored storage.

It’s a taste of things to come for the whole BBC and promotes collaboration and sharing of content cross genre, cross platform and supports the concept of 360 degree commissioning.
RETRAUPLOAD POLICY MANAGEMENT

Nearly all content related to programmes and online content is captured in the digital archive from 2007, with 100 percent of new Radio content and around 70 percent for new TV material. Legacy material is digitised on demand based on customer requests that come into our department by email and phone, and the requirement for a bespoke collection to be produced. We digitize about 300 archive programmes a month on demand.

Where previously we have been unable to capture all new Radio content we now capture it all digitally with no hard copy backup, and we capture all Television content on tape and make selections to digitise based on re-use and transmission criteria.

Digital production and archiving allows us to capture all content with a minimum metadata set attached, something we have never been able to do before. The addition of a minimum metadata set ensures that all

_Ill 1. Parallel Worlds, Parallel Lives._
content has basic information attached to allow if to be searchable and usable. At the point of cataloguing the Digital Library allows us to add retention policies according to BBC policies. Depending on the programme or item genre retention can be added automatically or manually, at any programme level and for any defined length of time with full reporting functionality to monitor the lifecycle of the asset.

The screenshot above shows an award winning network BBC programme Parallel Worlds, Parallel Lives, kept because it is highly requested. It will be kept permanently online, with tape back ups available. It’s catalogued with a documentaries genre. The genre being added alerts the system, a policy exists with the criteria of completed programme and documentaries. The system matches the programme and correct policy for Documentaries with retention to keep indefinitely. This policy is defined by the corporate media management policy.
This example above (ill 2) shows a News rushes tape that is retained for five years from the recording date. We defined through our BBC Scotland archive policy that this file has high re-use value but will date. We can assess the value of the content after 5 years has expired and the system will flag up the item needs reviewed. This policy was added manually, genre is not added through cataloguing, and instead the system allows an appropriate manual policy to be added from the current list that matches rushes with retention of five years, with review at the end.

You will notice that the policy does not have an automated disposal date. We decided to safeguard against automated disposal at this point and assess all content before disposal, with the possibility of extending the policy if the value had changed with the retention period. This might be if an interviewee died and these were only remaining pictures of them, or a building or event took on a new historical perspective.
The task flow management allows all polices to be monitored through the reporting tool and task alerts (ill 3). When a policy has expired and a review has taken place and disposal authorised, the task flow management passes on to the task of disposal. It is possible to stop disposal and add another policy at item level or proceed with no back up, or offer a physical back up (ill 4)

DIGITAL ARCHIVING SUCCESSES:

- Integrated digital workflow means the archive now captures all content produced.
- Delivering all new archive content online to users desktops
• Proactive archive management: delivering online content for complex playout, collection management, anticipating user re-use requirements

• All content cherry picked and ongoing process of digitising all content for re-use and preservation

• Dual running of legacy archive and online archive with decrease in content of tape as digital files increase

Where previously it was an aspiration in a tape based world, we now capture all content either on tape or digitally. We have started a process of tape elimination with digitising content on demand based on customer enquires, and proactive collection management to reflect upcoming events and anniversaries as well as production requirements. An example of this has been the 10th anniversary of Scottish devolution from the UK government. Various programmes and news coverage have been made to mark this event and the digital archive holds definitive collections of audio and video on the subject that have enabled the programmes to be made. This kind of proactive management of the archive has not previously been possible and the research and online representation has been done with the help of production colleagues to target their desktop programme making needs.

Due to the success of our digital working, we have attracted many new archive commissions into BBC Scotland, these have included landmark programmes such as the History of Scotland, a fully HD production. We have the ability to keep not only the final transmission programme but also the associated rushes. These rushes have since been used in other BBC programmes including another HD production A Portrait of Scotland. This collaboration would have been impossible with previous ways of working.
A digital workflow and archive has enabled BBC Scotland to launch a joint funded channel especially for Gaelic audiences, BBC Alba. Launched in 2008, the complex production and playout workflow requires a streamlined way of working that allows for transmission across the analogue and digital platforms with aggressive repeat schedules. Having a digital copy of both subtitled and un-subtitled material allows the productions and transmission flexibility and security with the programme. Rights management around this material is also handled through long and short term retention policies within the archive.

Finally the ability to reinforce archive resilience with mirrored storage on and offsite has given us more comfort than previously appreciated in a tape based world. The space and logistics issues associated with tape have left some areas of the archive vulnerable. However we now have added resilience in holding digital file copies of programmes. With ongoing preservation efforts this will also enable us to target areas of the archive that are becoming redundant such as D3 and Beta SP and SX formats.

DIGITAL ARCHIVING – UNEXPECTED CHALLENGES

- Dual running of legacy archive and online archive a long term process – Managing availability of useful online content vs. storage
- Storage – Costs high, moving to HD commissions; reporting and deletions not straightforward
- Outgrown corporate policies – aspirational, based on transmission material and linear workflows
- Resource – Digitising legacy content from tape and re-cataloguing content – resource heavy, prioritisation required.
• Technology – Automated Retention technology complex and still in infancy

We capture online 100 percent of Radio however some of the content is never referred to again. Talk radio and commercial music programmes have little re-use value and take much effort and resource to manage for very little value to our users. The question is we are meeting the media management policy to keep everything for five years post transmission, but if the content is not used at all, why are we putting the effort into managing it? The corporate policies have been difficult for Radio to achieve in the past with a network archive capture of only 30 percent of output. The dream was to achieve full capture and the policy was something to aspire to. Now we have achieved it in BBC Scotland but to what end? The cost both in storage and resource has made us question if the transmitted content needs to be kept online and catalogued, offline with no cataloguing or not at all in certain cases.

With the expense of online storage and the comparative ease of archiving on tape we will be running dual tape and online archives for some time yet. We budgeted closely for storage, but with the success of the digital working in Scotland we have gained many new commissions, launched a new channel and we are moving further into HD commissions. Being the first to go digital in the business, storage costs remain high in the BBC. We have also had difficulties with poor storage reporting tools, with media storage often treated in a similar way to documents and the preconception that if the storage fills up, the problem is easily solved by deleting content. Not ideal in a closely managed archive. Finally, if deletion is required at the end of a policy, it can be a cumbersome process across tiered and mirrored storage.

Is there more value in keeping the online content which has higher re-use than the transmission material, with the material of lesser re-use on tape? Or keep it online for a shorter retention period, with the corporate
media management policy being fulfilled through tape archiving? Radio presents a problem as its fully digital workflow. No back ups are produced. If we choose to dispose of some less used commercial music programmes or talk radio before archiving we would not be fulfilling the media management policy but we would be effectively using our resource and on-line storage and not necessarily letting down the user community.

There are no definitive answers to these new challenges, but it does pose the bigger question of why we have archive and retention policies in the first place. Is it that we would like to keep everything but have to make difficult decisions about what we can resource and store effectively, or is it we don’t believe there is value in keeping everything.

MEDIUM TO LONG TERM IMPRESSIONS

- Automated retention technology will improve to manage more complexity
- Storage costs will decrease as more of the BBC digitises, however this may be offset by increased tape elimination and HD and Super HD in the long term
- Linear and digital workflows will remain for some time and complex retention management will continue
- Polices will mature as experience increases and more areas of the BBC go digital

Very complex management of tape and digital retention will continue, but automation will improve. With the issues previously discussed around the complexities of digital storage and with a move in the BBC towards
public access to its archives it may be easier to keep everything. Polices will mature more around litigation, editorial censorship, ownership. This way of developing archive and retention would rely heavily on a decrease in storage costs, the reliance on minimum metadata sets and more automated retention capabilities within asset management systems with much less human intervention.
Henry Lindqvist has worked in the TV Broadcast business in YLE for more than 30 years. He established the videotape post production department in the late 1970’s and was head of that department in TV operations for 16 years. The video tape formats are since that time well known.

Since 1994 he has been working in the technical department as Project Manager in a great number of projects leading to the total digitized production and transmission environment as well as the tapeless workflow with file based deep archive solutions. Completed bigger projects in recent years are: Multichannel Transmission Automation, Enterprise Graphics system, Virtual Studio and now in process the Digitizing Center.

During the last decade he has been an active member and chair of the Media Management Commission in the FIAT/IFTA organization. He was chair of the Media Management Commission between 2000-2006 and member of EC 2002-2006.
ALREADY IN 2004, YLE started a large scale project, NASTA, in order to optimize all essential television production processes. The project documented and analyzed existing television production processes and work-flows in the major production categories.

From a technical point of view, the goal of the project was to integrate all existing and new digital production systems, as well as systems for the television archive, into a common networked production platform.

The core of this platform is called METRO, YLE’s enterprise-level media asset management system for all kinds of media. The METRO system administrates, transmits and archives in digital file format video-and audio material, subtitling files, stills and other data connected to the programme material as well as descriptive metadata.

In late 2007 YLE launched an investment project which target was to create and build a centralized Television digitizing center to canalize most
of the material to be digitized into the enterprise-level media asset management system METRO. The METRO is built around the Blue Order’s Media Archive concept. The digitizing center could be seen as the core of the whole Media Management system providing programme material with high capacity in file mode to the METRO.

Today’s situation is still quite old fashioned. Although the acquisition format in the news, sports and actualities area is P2 and edited as files on Avid workstations we are just in the point of starting to digitize in larger scale television programmes to the archive in file format.

THE PROJECT TARGET

In technical terms the project is divided and phased in many different workflows which all make several simultaneous signal transfers in file mode to the METRO system possible. The purpose of the project is also to establish a most effective and industrial digitization process.

The center should:

1. produce high quality digital video material with connected metadata information into METRO to be accessed by the programme production and scheduling department and transmission automation.

2. supply video material to be accessed by all programme production personal via METRO. The target is to provide better and simultaneous accessibility to a broader usergroup than today independent of time and place.
3. support the tapeless or less tape programme production bringing savings in videotape costs and in videotape storage-, logistics-, copying- and migration costs.

4. make the video material long-term storage and accessibility possible for decades ahead.

5. promote completely new ways of using video material

The project limitations are:

1. the center does not in the starting phase support our AVID production environment. Acquisition material will be stored on disc or on memory card and transferred into our I-Play system from decentralized drivers. The produced final programme material will be transferred to METRO directly in file mode.

2. the center will not take care of fastly repeated programme transmissions. Live transmissions that will be transmitted on another channel very soon after the first transmission will be handled in our production environment. Sometimes the second transmission starts before the first transmission has ended. This requires some editing procedures which do not so far belong to the Centralized Ingest Area working fields.

3. the center does not have responsibility of “delayed” transmissions. This represents transmission cases where the beginning of the recorded programme goes on air simultaneously as a later part of the programme is recorded at the same time and some times changes to live transmission during recording.
The Project Outline/The Central Ingest Area incorporate the following digitizing processes:

1. A storage buffer and recording facilities for live transmissions to be converted and archived in file format. Includes back-up capacity.

2. Digitizing of programme material on to the transmission server and the media archive.

3. Digitizing of material coming from the international programme exchange for local editing.

4. On-demand digitizing of archive material for programme production:
   - automated bulk digitizing, the system will use existing Betacart machines for this
   - editing required material (connecting, splitting, cleaning and other technical corrections)

5. Rush digitizing service of insert material to programme production

6. Case by case or multiformat videotape transfers for our own production

7. Not officially in-house supported file format programmes conversion to the METRO supported file format
The facilities and functions in the Centralized ingest area are divided into several independently working suites including different working functions. All are striving to the same goal: digitizing programme material to the METRO is in a unified file format. The quality control of the programme material going through this digitizing center is playing a great part in the working process. The quality of the programmes are being checked in several phases and the parameters are being changed if needed so that all material we transfer to the METRO Is in a unified format. This means in practice that when we retrieve and reuse material from the METRO we should not any longer need to do quality checks. We can move the material for instance directly to the transmission server for transmission without control. When the system is completed and implemented to the planned
configuration and all workflows are running we should be able to transfer 200 hours of programme material per day to the METRO. This means in standard definition quality (50Mbit/s) that the media archive will grow with around 10PB per year the following 10-15 years.

Before this digitizing of old legacy material is finished there is already a migration need for the material we started with to be digitized to newer storing carriers. This is becoming a never ending transferring cycle. How many television programmes are we willing to store on one single carrier? It is possible today to store 100-200 psc. of average length programmes on one carrier. Is it more secure to have less programmes per carrier than as many as the technology allow?

We have put special efforts on the quality control of the material. Although we have to permit a certain amount of wastage due to the enormous amount of material. Due to personal capacity there is no possibility
to do a real time checking of everything. Only spot checking is being done. Anyway we know that all legacy material have sometime been produced carefully with several technical quality control steps during acquisition, editing and transmitting. If we find bad tape batches we naturally process the control of these with special care.

Normally we have a mechanical check of the material in the input stage, we set the parameters for the video and audio levels and check how the audio tracks are used. When the programmes are in file mode in the central server we spot check the files before we deliver the programmes to the media archive.

**A NEW WORKFLOW**

Apart from this fairly huge technical deployment and new opening this project is also much about implementing new workflows and integrating different organisational working processes. We will change the working character and process in the existing tape recording and transferring center. The working process and the job descriptions will be totally new. We proceed in stages by training small groups in accordance with the time table we take new technical units in operation. This is part of the natural developing process when we move into a less tape working environment.

The whole Centralized Ingest Area will stand under our TV Archives administration. From organisational point of view we see a movement of technical staff working in the archive in the near future. To get the staff more motivated in the handling of this process we give the operators responsibility of the whole transferring change. This allows of a more flexible use of our resources. One operator fetches the tapes from the deep archive, takes care of the transferring, quality check and the return of material.
THE CENTRAL INGEST AREA’S WORKING UNITS

1. Videotape forwarding or expediting
   
   • This is the mechanical identification, checking and forwarding of the videotapes from the tape archive to the digitizing area.

2. The Preparation unit
   
   • In this unit the tape cleaning is executed and all necessary information from the tapes are being checked and stored into the TROLLI application. The TROLLI is an “in-house” software application which will be described later on. This information will then be changed into control orders for the tape machines and other necessary devices.

The parameters that will be checked and stored are:

   • Video and audio levels
   • The start and end point of the recording
   • Use of the audio tracks
   • Aspect ratio
   • The programme content information
3. The Flexicart Unit

- In this unit are all the Beta family tapes mass digitizing into the media archive taking place. We have three Flexicart machines with three digibeta machines in each robot. The control and administration of the machines is taken care of by the YLE “in-house” software application TROLLI.

- The Flexicarts reads the barcodes on the cassettes into its database.

- The Trolli application gathers the information to be used as control of the recorders.
4. The Flexicart Control Units

- These units take care of the control and transfer of the machinery, a quality check of the playback process and a post check of the generated files.

These units take care of the following functions:

- During playback a Waveform monitor store the video and audio levels alarm values exceeds

- The operator check the programmes due to alarms and also randomly
• If the recorded material does not fill the quality criteria is goes to manual transfer

• On the base of the operators approval confirmation the recorded files will be transferred to the Media Archive

5. 1” C Migration Unit

• The 1” C material (30,000 pcs tapes) is being migrated in a unit especially designed for this. A semi automated control software application created “in-house” is in use. In the beginning we will digitize from two 1 inch recorders, later on we will expand this amount. This unit has been in
operation now for three years and we have so far migrated some 10,000 tapes to Digibeta. These digibeta tapes will later on and on-demand be digitized in the Flexicarts.

6. Unit for converting and editing files coming from external sources
   - The Centralized Ingest Area will have a unit for controlling and converting files coming from external programme sources which differs from our standards.

7. The T5 Unit
   - This unit will take care of all non Beta family tapes migration. The control system and quality checking procedures are the same as with Beta tapes but the playback is done with single recorders. No robotics are used. DVCAM and DVCPro formats are being digitized here.

8. Recording of live feeds
   - One very important and big working field for the Centralized Ingest Area is to take care of the recording and digitizing of the live feeds. The center is responsible for recording live feeds from all our studios, OB’s, national and international programme sources. During the SD era these programmes will be recorded in IMX 50 format. A Secondary Event function in the transmission automation trigger the Centralized Ingest Area server to do these recordings.
9. Tape editing and digitizing Unit

- This is a special unit for quality checking of problematic video tapes and editing.

- The unit also splits videotapes containing of several programmes into separate files as well as gluing together programmes recorded on several tapes due to the length for instance into one file.

- Audio levels are being corrected here as well as swapping of audio channels up to standardized recommendations.

Digitizing live transmissions, YLE Centralized Ingest Area.
10. Tampere

- I the regional center in Tampere we will build a similar tape digitizing unit as the T 5 in Helsinki.

- The Tampere archive holds some 13,000 pcs of D3 tapes. The transfer of these tapes are getting time critical due to the almost non existing support of the format. A D3 automated robot has been built to speed up the migration of these tapes.

- Material digitized in Tampere will be transferred to the Media Archive over the net.

**DESCRIPTION OF THE TROLLI SYSTEM**

The TROLLI software is a YLE “In-house” made application which basically is an administration and control software. TROLLI administrate the video tape originated material’s digitizing process all the way from that the programme order has been generated in the METRO until the updated programme information with the digitized programme is being sent back to the METRO.

The TROLLI system consists of three main applications:

1. Order management

   - The order management receives a programme digitize order from the METRO and administrates the order until it can be returned to the METRO
2. Check Point

- The METRO order’s programme related information is being checked from tape. Checked information is for instance the programmes start and stop times, audio and video levels.

3. Capture Control

- Control of digitizing equipment. Three Flexicart and one ProCart tape robots are used. Besides the tape robots Capture Control administrate Ingest and digitizing equipment, video and audio on-line quality checking and audio tracks swapping equipment.

The TROLLI system consists of the following secondary applications:

- Trolli Live

- The application reads the secondary event function from the transmission automation and triggers the Centralized Ingest Area server to do live transmission recordings.

- Final Check

- The file based error messages are being checked. The final Check is the unit where The original METRO order will be returned to the METRO.
• Edit
  • Final Cut Pro is used as the editor because it can see the material on the Omneon server directly on the timeline. Programmes stored on several different tapes will be glued together into one file. If necessary small cosmetic changes will be done to the programmes.

• Trolli Main Monitor
  • A real time display of all the activities in the Centralized Ingest Area

• Trolli File
  • File converting and checking. Incoming files which do not correspond to the METRO standard will converted to the agreed format.

THE PROJECT'S TIME SCHEDULE

As mentioned earlier the project will be implemented in several phases starting this year. The target is to finalize the project in 2010. The first testing equipment has been installed. Things look good and the first unofficial transfers to the METRO has been done.

Phase one which incorporate digitizing of live transmissions is to be taken in operation in September this year. After this we start with the process linked migration of all different legacy material in steps that from workflow and administrator point of view are controllable. Our target is to get 100 hours per day through the Centralized Ingest Area this year and another 100 hours next year so that the final throughput will be 200 hours per day.
YLE CENTRALIZED INGEST AREA
Gerhard Stanz studied communication science at the university of Vienna and works in the archive of the ORF since 1991. Technology-Implementations – mainly in the area of preview – and European research projects led him to his current position as “systems developer” in the television archives. With focus on content management and metadata, he was involved in several projects of the ORF like the modernization of the newsroom, the introduction of digital production in the regional studios and lately the implementation of the new archive-system “FESAD”. His current activities are the build up of a digital archive and the integration of the archive into the file-based production workflows at his company.
ALL DECISIONS REGARDING retention can only be based on metadata; therefore a retention policy also requires a corresponding metadata policy. Almost every “area” of metadata, can contribute to qualified retention decisions.

Without going into detail on the level of individual fields / attributes, I try to give an overview which decision criteria can be derived out of the respective metadata areas:

CONTENT-METADATA / DESCRIPTIVE METADATA:

Any Idea of the subject matter and the actual picture content helps to estimate the re-use value, which in case of a selection is the focal point for retentions decisions. This is especially the case for rushes and raw material.
RIGHTS-/LICENSE METADATA:

For broadcasters and other commercial users of video-material, the availability of licenses is one of the most important factors that determines the re-use value of a given content and therefore rules retention in this area.

TECHNICAL METADATA:

Technical Metadata might be formal information about the media (be it a file or a tape) which influences comparative retention (when one has the same content on several media, the technically “richest” format would act as be the master-copy for migration which determines the other copies for deletion. A classical decision in a retention process.) Technical Metadata could also be the result of a quality-check.

USAGE METADATA:

This can be metadata documenting how often a piece of content has been ordered from the archive, but of course that also can be metadata that shows out of which clips a programme is made up.

In the tape world, the most used content would e.g. be qualified for migration, whereas for the most used media, the opposite would be the case.

In general one might say, that there can never be enough metadata as a basis for retention decisions even so technical copy histories or usage metadata might look confusing on first sight. The challenge is to properly structure the data in order to utilize it for retention.
The key factor to improve metadata management is integration of metadata systems. All file based production is more or less database driven and the content should enter the archive with as much accompanying metadata as possible.

E.g. tracking the source information (which soon will be a standard feature of each CMS/MAM) will help to take the relation between raw material and finished programme more efficiently into consideration. The copyright status of edited material will be easier to track and can therefore better act as a basis for retention decisions.

Basically, we are – from a situation where material is often to be deleted because of a lack of metadata – on the way towards a situation where deletion will more often take place based on abundant metadata. Hopefully, this will lead to better decisions in deleting the obsolete in order to concentrate resources on the sustainably useable.
As a project manager at the French Institut National de l’Audiovisuel, Ina, Dominique Saintville is presently in charge of international cooperation projects regarding media archives. She coordinates the Ina Sup course “Expertise des fonds” (assessing archive holdings). She has been working in media archives since 1977 with information systems and metadata methodology, marketing, preservation and digitisation programme, history of TV programmes and archives. She has also worked for the Fondation nationale des sciences politiques in Paris and the UNESCO in Niger. She has been involved in FIAT/IFTA activities since its creation and is presently senior vice president and a member of the Programme and Production Commission. Her academic background is qualified in political science (Sciences-Po, 1966) Ph. D in information & communication science (EHESS, 1979).
THE ISSUE OF selection has always been at the heart of all archive policies – rules should be established for collecting new material and for preserving the archive in the long term. The digital revolution has created a new context: the analogical material are becoming obsolete whereas the demand for audiovisual content to be used on digital networks is growing fast.

The urgent need to digitize the legacy archive has set a new challenge to archive managers. How to deal with the bulk of old film and videos? Should they be all saved or should priorities be defined? To what cost and to what benefits?

Each company has its own rationale, depending on its missions and strategic options. Inas main mission is the preservation and exploitation of the national broadcast heritage. The decision was taken in 1999 to start a massive preservation and digitization plan. Soon after its start, the issue of selection had to be addressed, and in 2003, it was decided to save all archives at risk, i.e.
some 835,000 hours of radio and TV programmes. The plan is due to last until 2015 and the total cost has been evaluated to 193 million euros.

This chapter presents the Ina experience. The historical approach underlines the way issues have been addressed and solutions have been found.
AU TOURNANT DU 20ème siècle, tous les centres d’archives audiovisuelles se sont trouvés confrontés aux mêmes questions touchant le devenir de leur patrimoine : fragilité des supports, changements des formats techniques, obsolescence des machines de lecture, volumétrie importante compte tenu de l’âge désormais respectable des chaînes de radio et de télévision publiques, etc. De plus, la numérisation des activités de diffusion et de production, la nécessité de pouvoir exploiter aisément un patrimoine né analogique, les perspectives d’accès ouvertes par le numérique et les réseaux, ont précipité la recherche de solutions adéquates de préservation des contenus archivés, et la conception d’ambitieux plans de sauvegarde et de numérisation.

Chaque institution a apporté des réponses, souvent similaires sur le plan purement technique, mais très diversifiées dans leur mise en œuvre : il s’est agi en effet de tenir compte des missions propres à chaque institution (diffuseur, archive nationale), des cultures (gens de télévision et de radio, gens de cinéma), des priorités et objectifs stratégiques.
L’INA, UNE INSTITUTION SINGULIÈRE

L’Institut National de l’Audiovisuel ne fait pas exception. Son plan de sauvegarde et de numérisation, lancé en 1999, fait aujourd’hui figure de modèle, ne serait-ce que par l’ambition qui lui a été donnée, quelques années plus tard, de traiter d’ici 2015 l’intégralité des fonds analogiques « menacés », soit un million d’heures de programmes de radio et de télévision. Cet objectif qui représente un effort continu sur dix sept années, est forcément lié à la stratégie mise en œuvre par l’Institut et à la spécificité de l’entreprise Ina.

En effet, l’Ina, établissement public à caractère industriel et commercial, né en 1974 de l’éclatement de l’ORTF, occupe une place singulière dans le monde des archives, se positionnant à la fois dans le domaine des médias, comme le sont la BBC ou la RAI, et dans le domaine culturel, à l’instar des grandes institutions patrimoniales que sont la Bibliothèque du Congrès, la BNF ou les Archives nationales. Cette double appartenance trouve ses fondements dans la dualité même du système juridique qui encadre son activité : la loi sur la communication audiovisuelle, la loi sur le dépôt légal. Elle se retrouve dans la nature des fonds conservés : le fonds des archives « professionnelles », qui remonte aux origines de la radio et de la télévision et qui s’adresse aux professionnels ; le fonds du dépôt légal, plus récent et plus étendu qui s’adresse aux étudiants et aux chercheurs.

LE PSN AUJOURD’HUI

Dix ans après le lancement du PSN, les résultats sont là et les chiffres sont éloquents : 360.000 heures de télévision et 180.000 heures de radio ont été sauvegardées et numérisées. 540.000 heures de programmes sont donc accessibles en ligne par les professionnels ; une large sélection de 22.000 heures correspondant à 100.000 documents est offerte en consultation au grand public.
Dix ans après le lancement du PSN, les procédures sont établies, de nouveaux outils ont été développés, le système est rodé, industrialisé; de nouvelles versions des services d’accès sont sur le point d’être déployées. Ce qui apparaît aujourd’hui comme une évidence, pour les praticiens comme pour les usagers de l’archive, est le produit d’un cheminement de dix ans. Mais se souvient-on encore des commencements ?

J’ai eu le privilège d’être associée aux équipes chargées du lancement et de la mise en œuvre du Plan de sauvegarde et de Numérisation (PSN).

Ce sont les Archives « professionnelles » qui sont concernées par le Plan de Sauvegarde et de Numérisation (PSN).

Dix ans après le lancement du PSN, les procédures sont établies, de nouveaux outils ont été développés, le système est rodé, industrialisé; de nouvelles versions des services d’accès sont sur le point d’être déployées. Ce qui apparaît aujourd’hui comme une évidence, pour les praticiens comme pour les usagers de l’archive, est le produit d’un cheminement de dix ans. Mais se souvient-on encore des commencements ?

J’ai eu le privilège d’être associée aux équipes chargées du lancement et de la mise en œuvre du Plan de sauvegarde et de numérisation. J’ai relu les notes, les études, les rapports qui documentent l’histoire de ce grand chantier. C’est par ce biais que je souhaite évoquer ici le PSN : dire les évidences, la constance de la volonté politique, mais aussi les difficultés, les interrogations, les évolutions ; montrer comment le chemin a été trouvé, comment les choix ont été infléchis ; partager cette expérience exceptionnelle avec ceux qui, comme l’Ina, ont déjà fait un grand bout du chemin,
comme avec ceux qui, aujourd’hui encore, le cherchent.

Dans cette perspective, on rappellera d’abord le contexte dans lequel a été lancé le PSN ; puis on passera en revue les grandes questions qui se sont posées et les réponses qui ont été apportées. Enfin, sur la base de quelques exemples, on montrera comment l’Ina a organisé sur ce terrain une coopération fructueuse tant avec ses partenaires européens qu’avec ses partenaires plus lointains, en Afghanistan ou en Afrique.

RETOUR VERS LES COMMENCEMENTS

Depuis toujours, la politique d’archivage de l’Ina a incorporé en tâche de fond des opérations de restauration et de transfert des programmes sur des supports « modernes », sinon pérennes. Mais le rythme de transfert, qui pourtant s’était accéléré en 1997 pour atteindre quelques milliers d’heures annuelles, s’est avéré très insuffisant pour faire face à la dégradation des supports film et bandes magnétiques, à l’obsolescence des formats vidéo, aux volumes à sauver.

La situation devenait critique, générant des lenteurs dans l’exploitation, et un mécontentement des usagers qui attendaient de l’Ina un meilleur niveau de service. Le délai de traitement des commandes de production atteignait alors couramment trois semaines !

PREMIERS PAS

Cependant, la révolution technologique en cours imposait un nouveau défi, le passage de l’analogique au numérique. Jusque là, le numérique à l’Ina, était principalement le territoire du service de la Recherche. Aux Archives, la maquette d’une banque de données numériques interrogeables à distance avait été montée dans le cadre des “autoroutes de l’information”;
en 1998, fut mis en service la première application numérique opérationnelle, baptisée AGPE, un système de numérisation des extraits vidéo, résultant des repérages effectués par les clients, et permettant de capitaliser ces extraits en vue de leur réutilisation.

L’INA FAIT LE CHOIX DU NUMÉRIQUE

En 1999, l’Ina définissait une stratégie d’entreprise fondée sur le numérique. Pour les Archives, cette stratégie s’est traduite dans le lancement conjugué du Plan de Sauvegarde et de Numérisation et du Plan Patrimoine.

Il n’est pas inutile de s’attarder quelques instants sur les objectifs de ce Plan Patrimoine qui a engagé la modernisation radicale des Archives. Il s’agissait, d’adapter les archives aux nouvelles pratiques d’exploitation des images et des sons et aux changements technologiques induits par le numérique, en créant une collection numérique de grande amplitude, et en développant les outils permettant de gérer et d’exploiter cette collection.

Pour ce faire, plusieurs chantiers ont été ouverts, parmi lesquels :

• des chantiers techniques : mise en place d’une infrastructure de stockage de masse, développement d’un système de workflow permettant de fluidifier les transactions liées au traitement d’une commande

• des chantiers documentaires : normalisation du traitement documentaire, afin de faciliter les échanges de données, celles venant des diffuseurs et celles qui circulaient entre les différents secteurs de l’Ina travaillant sur les archives ; développement d’un système documentaire de nouvelle génération, permettant d’accéder en ligne aux documents numérisés et de structurer les fonds (segment-
tation des programmes, documentation des segments, création de dossiers thématiques), baptisé ultérieurement TOTEM

- des chantiers juridiques : négociation d’accords collectifs avec les différentes catégories d’ayants droit en vue de faciliter l’exploitation des archives sur les nouveaux réseaux ; numérisation des dossiers de production

Le Plan de sauvegarde et de numérisation constituait la pierre angulaire de ce plan Patrimoine. Tel que défini en 1999, le PSN avait pour but:

- de sauvegarder le patrimoine audiovisuel « en danger », évalué à 220.000 heures, en créant des copies de préservation

- d’offrir un accès élargi aux contenus en numérisant « une grande partie » des archives

- et ainsi, de permettre une exploitation des fonds à la fois plus efficace, pour offrir un meilleur niveau de service à nos partenaires institutionnels que sont les diffuseurs publics, et plus étendue, la numérisation des fonds devant permettre de développer sur une grande échelle la commercialisation des archives.

L’INFRASTRUCTURE TECHNIQUE DU PSN

Les choix des formats numériques ont été effectués en fonction des technologies disponibles en 1999. Il s’est agi du Béta numérique pour la copie de préservation, des fichiers MPEG 2 à 8 Mb/s pour l’exploitation, des fichiers MPEG1 à 1 Mb/s pour la consultation en ligne. On considère aujourd’hui de nouveaux formats, comme le H264 à environ 400 Kb/s.
pour la consultation en ligne ; et des formats de fichiers pas ou peu compressés pour la copie de préservation (en remplacement de la Béta numérique), comme le MJPEG 2000 ou le H264 intra.

Les chaînes de transfert ont été conçues en fonction de la nature et de la qualité technique des supports et enregistrements d’origine. Ainsi pour le film d’actualité, les sujets d’une même journée font l’objet d’une remise en état mécanique, puis ils sont mis bout à bout pour passer au téléséminé, donnant naissance à une Béta numérique qui sera encodée en MPEG 1 et MPEG 2. Les cassettes ¾ pouce, supports des programmes de flux (parallèles antenne), sont nettoyées, et leur contenu est encodé directement en MPEG 1 et MPEG 2, à la différence des bandes vidéo 2 pouces, supports des programmes de stock (fictions, variétés), qui sont transférées sur un master de conservation Béta numérique avant encodage.

L’objectif de numérisation fixé par le premier Contrat d’objectifs et de moyens pour les années 2000 à 2003, a été établi à 123.000 heures de télévision (et 40.000 heures de radio). Il a été atteint grâce à la mise en œuvre de trois ensembles de systèmes techniques :

- le premier dispositif, qui a pu démarrer dès 1999, est constitué des chaînes de traitement externes, gérées pour le compte de l’Ina par des prestataires

- le second dispositif, interne celui là, appelé SNC (Sauvegarde, Numérisation, Communication), est dédié en priorité aux travaux de sauvegarde et de numérisation liés aux demandes des clients. Les chaînes SNC ont été opérationnelles en janvier 2001

- Le troisième dispositif consiste à capter en numérique la programmation courante des chaînes. Il a été mis en œuvre par l’Inathèque dans le cadre de sa mission de collecte des nouveaux programmes. Ce dispositif a fonctionné pour la télévision à partir de 2001
Tels furent les commencements du Plan de Sauvegarde et de Numérisation.

**LES GRANDES QUESTIONS**

L'on avait fait les choix stratégiques et techniques. L'infrastructure était en place. Les premiers résultats étaient tangibles. Mais ce n’est qu’après le lancement du PSN que l’on prit une pleine conscience des questions qu’il restait à instruire, et pas des moindres : la volumétrie, la sélection, le financement, etc.

Cette période de questionnement débute au moment même où Emmanuel Hoog est nommé à la présidence de l’Ina, en 2001. Son arrivée va permettre de consolider l’édifice encore fragile du PSN, de l’installer dans la durée, de lui donner une ambition nouvelle, et de construire, à partir d’un fonds numérisé de plus en plus conséquent, une galaxie de services d’accès à ce patrimoine.

Avec lui, l’Ina va s’engager dans la sauvegarde et la numérisation de « toute » la mémoire audiovisuelle archivée, et dans la communication de ce patrimoine. Pour réaliser cette nouvelle ambition, il a fallu instruire un certain nombre de questions essentielles que nous allons passer en revue.

**LA VOLUMÉTRIE**

Une première évaluation des matériels de télévision et de radio à sauvegarder avait été réalisée en 1999.

- Pour la télévision, on avait décompté 220.000 heures de programmes, représentant 51% des fonds archivés : soit, 220 h film nitrate, 56.000 h film, 22.000 h de bandes vidéo 2 pouces, 12.000 h de bandes vidéo 1 ponce B, 130.000 h de cassettes Umatic
On notera que les bandes vidéo 1 pouce C et les cassettes Béta SP analogiques, n’étaient pas alors considérées comme matériels en danger. L’on est revenu depuis sur cette appréciation.

Pour la radio, on avait décompté 305.000 heures de programmes, représentant 52% des fonds : soit, 20.000 h sur disques 78 tours à gravure directe, 280.000 h de bandes magnétiques, 5.000 h de cassettes DAT.

Face aux difficultés rencontrées au cours des premières années, et constatant le rythme d’avancement des travaux, il était clair qu’une partie importante du fonds serait sacrifiée si l’on n’accélérerait pas le plan. En effet, au rythme d’alors, et considérant la durée de vie des supports archivés, l’Ina n’aurait pu sauvegarder que 37% du fonds télévision et 9% du fonds radio!

C’est avec le double objectif de sauvegarder « tout » le fonds, et d’accélérer le PSN qu’un audit du plan de sauvegarde fut confié à la société Veritas. Il s’agissait d’évaluer les volumes de matériels archivés restant à sauvegarder, le rythme nécessaire pour tenir compte de la durée de vie des supports les plus fragiles, et les coûts afférents. L’Ina devait en effet s’appuyer sur une évaluation fiable pour évaluer les coûts et trouver les moyens financiers nécessaires à la mise en œuvre d’un plan intégral et accéléré.

Les résultats de l’étude furent communiqués en janvier 2003:

Les évaluations des matériels en danger furent relevées, pour atteindre respectivement 336.000 heures de programmes de télévision et 498.000 h de radio, soit un total de 835.000 heures.
(ramenées à 740.000 heures, si l’on tient compte des 95.000 heures sauvegardées au cours des années 1999 à 2002)

• L’échéance de fin du PSN est fixée à 2015

• Le coût total du PSN est évalué pour les années 1999 à 2015 à 193 millions d’euros.

Pour atteindre ces nouveaux objectifs, l’audit a formulé un certain nombre de recommandations visant notamment à:

• Renforcer les équipes Ina en charge de la gestion du PSN et à les doter d’un outil informatique de suivi des opérations. Aujourd’hui, soixante dix collaborateurs Ina sont affectés à la mise en œuvre du PSN. Ils travaillent en liaison avec trente prestataires, mobilisant cent cinquante personnes affectées à des tâches de transfert technique et d’inventaires.

• Organiser la programmation systématique du PSN, en affichant des objectifs précis et argumentés de sauvegarde et de numérisation, année après année

• Développer des méthodes de travail industrielles (chaînes de transfert « parallélisées »)

• Concevoir un système de stockage de masse adaptés à la volumétrie. Aujourd’hui, les Archives disposent d’un système de stockage de masse évolutif, permettant d’accueillir la totalité des programmes numérisés, radio et télévision – sur disque dur pour un accès en ligne et sur des bandes LTO3 gérées dans une librairie robotisée ADIC
TOUT SAUVEGARDER?

La première réponse à la question incontournable de la sélection a été technique. On a commencé par sauvegarder les matériels les plus vulnérables : les films nitrate des années 40 et 50, les parties identifiées du fonds film atteintes du syndrome du vinaigre, les programmes enregistrés sur un support unique (film ou vidéo), les formats vidéo obsolètes (bandes 2 pouces). En parallèle, on a donné la priorité aux collections les plus demandées : les journaux, les magazines d’information, les émissions de variétés.

Au-delà de ces choix évidents, des questions plus ardues se sont posées:

- Comment aborder la masse des émissions archivées sur plusieurs matériels (que ce soit du fait des techniques de production, ou parce qu’une copie d’exploitation avait été réalisée). Quel matériel choisir ? Comment identifier les doublons ?

- Comment arbitrer des choix de contenu ? Quels critère privilégier : l’intérêt culturel, historique, commercial, scientifique ?

- Comment gérer l’écart colossal entre les volumes en jeu, la capacité limitée de traitement dans le temps, les ressources financières disponibles ?

Il fallait échelonner les travaux. Il devenait urgent de définir une méthode permettant de décider quelle émission, sur quel support, envoyer en sauvegarde, et à quel moment. Et pour établir la méthode sur des bases fiables, il était tout aussi impératif d’améliorer la connaissance que nous avions de nos fonds, tant sur le plan des contenus que sur le plan technique.

Un projet Connaissance des fonds avait été lancé en 2001 pour accompagner la fusion des équipes travaillant sur les fonds de l’actualité et de la
production télévisés. Ce projet a permis de donner une vue d’ensemble des collections archivées, sous l’angle de l’histoire des programmes, une vue plus qualitative que la vue très analytique fournie par les bases de données documentaires. Ce travail, qui a mobilisé de nombreux collaborateurs des archives, chacun détenant un savoir précieux sur telle ou telle partie du fonds, a débouché sur des stages de sensibilisation des personnels et sur l’édition de dossiers décrivant les grandes catégories de programmes (le journal télévisé, les fictions, les variétés, les documentaires, les sports), un fonds particulier (la presse filmée), ou une problématique essentielle pour la compréhension des archives (l’évolution du traitement documentaire, l’histoire des supports, la gestion des droits).

C’est la même année, à partir du mois de juin, que des campagnes systématiques ont été lancées pour évaluer le degré d’atteinte du fonds par le « syndrome du vinaigre ». Nous savions qu’une partie du fonds film et des bandes radio était « vinaigrées », mais nous n’en avions pas la cartographie : quelles collections étaient touchées, sur quelles années, quel était le niveau et le rythme de dégradation, où fallait-il intervenir immédiatement ?

Pour nous accompagner dans la définition des priorités, un Comité de sauvegarde, présidé par Jean-Noël Jeanneney, composé d’une douzaine de personnalités extérieures, représentant la communauté des créateurs et utilisateurs d’archives, a été mis en place. Il a été chargé de définir avec les responsables Ina du PSN, les parties du fonds à « sauvegarder en priorité ». On notera que le terme de sélection avait été banni ! Cinq séances, programmées de mars 2001 à juin 2002, ont été consacrées respectivement aux émissions d’information, aux magazines et documentaires, aux fictions, aux émissions musicales, au fonds « parlé » radiophonique. Elles ont contribué de manière significative à l’enrichissement et à la clarification des problématiques de sauvegarde.

Tous ces travaux, relatifs à la volumétrie et à la sélection, ont débouché fin 2003 sur une redéfinition du plan de sauvegarde, conformément aux préconisations de l’audit Veritas, et sur l’élaboration d’une grille de
programmation complète et sophistiquée de réalisation du plan “accéléré”, prenant en compte l’ensemble des paramètres – techniques, catégories de programmes, collections, volumes horaires, coût, échelonnement dans le temps – pour dire précisément dans quel ordre et suivant quelles logiques traiter les collections, et ce jusqu’en 2015.

FINANCER UN PSN ACCÉLÉRÉ

Le coût des opérations de sauvegarde et de numérisation reste élevé même si la mise en place de chaînes industrielles et le caractère massif du plan ont permis de réduire très sensiblement les coûts unitaires des prestations, par rapport aux prix usuels pratiqués sur le marché. A cet égard, il convient de noter que le traitement du film est infiniment plus onéreux que celui de la vidéo. La sauvegarde d’une heure de film coûte de 6 à 10 fois plus cher que celle d’une heure de vidéo. La chaîne de traitement film inclut en effet des opérations de remise en état mécanique de la pellicule, de montage, d’étalonnage, qui sont autant d’opérations manuelles minutieuses et longues.

Le coût total du PSN accéléré était évalué à 193 M€ : comment a-t-il été financé ?


Dès lors qu’il s’est agi de passer à la vitesse supérieure, l’Ina a dû convaincre les pouvoirs publics, les parlementaires, les professionnels et l’opinion
du bien fondé d’un plan de sauvegarde et de numérisation “intégral et accéléré”.

On a évoqué plus haut le Comité de sauvegarde qui a su exprimer avec éloquence la valeur patrimoniale, artistique, professionnelle, commerciale, scientifique des archives, et l’impossibilité de sélectionner, car cela voudrait dire que l’on serait prêt à abandonner des pans entiers du patrimoine. On a évoqué aussi le rôle clé joué par les résultats de l’audit Veritas qui a su chiffrer l’ambition du projet.


Il faut aussi rappeler le colloque organisé le 1er avril 2003 par l’Ina à l’Assemblée nationale « Les archives de la radio et de la télévision : un nouveau patrimoine à sauvegarder ». Ce fut l’occasion d’expliquer publiquement les menaces qui pèsent sur les archives audiovisuelles, de souligner les enjeux attachés à la préservation du patrimoine audiovisuel, de répondre par anticipation aux questions que soulève la mise en œuvre d’un plan accéléré, de montrer que la BBC et la RAI avaient entrepris des projets de même ampleur.


Les ambitions portées par l’Ina et la constance de ses efforts pour la sauvegarde du patrimoine audiovisuel national étaient désormais reconnues et validées. La consolidation de la dotation de redevance ouverte en 2005 et les compléments prévus les années suivantes impulseront, dès
2006, une nouvelle accélération du PSN pour le porter à un niveau qui permettra la sauvegarde intégrale des collections menacées à l’horizon de 2015, conformément au scénario de sauvegarde fondé sur les conclusions du rapport Véritas.

Mars 2009, nous arrivons bientôt au terme du COM 2. Les engagements ont été tenus, plus de la moitié des fonds menacés est désormais préservée ce qui devrait permettra à la France d’être le premier pays à pouvoir sauver son patrimoine radiophonique et télévisuel. Le COM 3, actuellement en gestation, devrait prendre en compte la suite du PSN et la sauvegarde du fonds Béta analogique, laissé de côté lors de l’étude Véritas. C’est un fonds qui vieillit mal et que l’on doit considérer aujourd’hui comme menacé.

LES NOUVEAUX USAGES DU NUMÉRIQUE

Il va sans dire que le plan de sauvegarde et de numérisation de l’Ina n’a de sens que s’il débouche sur un élargissement de l’accès au patrimoine de la radio et de la télévision. C’est d’ailleurs une autre des orientations fondamentales inscrites dans le COM 2 : rendre accessibles à tous les publics, y compris au grand public, les collections audiovisuelles gérées par l’Ina. La diversité des publics et des attentes a été prise en compte.

Le service Inamediapro.com, lancé en 2004, s’adresse aux professionnels des médias, français et étrangers. Il donne accès à l’ensemble des archives numérisées. Dès lors qu’il est accrédité, l’usager peut conduire lui-même sa recherche (recherche multicritères dans une base de données, recherche guidée dans une arborescence thématique), consulter en ligne les vidéos, poser des repères et des commentaires, constituer un panier, et l’envoyer au service commercial pour analyse des droits, devis et recopie.

Dès 2005, le chantier pour la constitution d’une offre accessible au grand public “Archives pour tous” était lancé. En avril 2006, le service
était ouvert sur le site Internet de l’Ina, avec un succès retentissant, qui ne s’est pas démenti depuis. Le service « Archives pour tous » donne accès à quelques 100,000 documents à travers une interface « éditorialisée », renouvelée chaque jour, organisée en rubrique, à l’image d’une chaîne web. Il enregistre 1,000,000 de visiteurs uniques chaque mois.

Il faudrait aussi mentionner les produits élaborés notamment pour le secteur éducatif, comme les « Jalons pour l’histoire du temps présent » qui proposent des collections plus restreintes, accessibles en ligne, sur des thèmes en rapport avec les programmes scolaires, un appareil pédagogique très poussé : fresques, accès chronologiques et thématiques, transcriptions, analyses contextuelles.

Ces services s’appuient sur le travail conjugué des documentalistes des Archives et des rédacteurs du secteur de l’Edition, ouvrant chacun avec des outils dédiés à ce que nous appelons la « thématisation », la valorisation et « l’éditorialisation » des fonds. Il s’agit non seulement de documenter les programmes, mais aussi de les fragmenter, de les organiser par thème, par personnalité, de les rééditer pour diffusion sur de nouveaux vecteurs (téléphones mobiles), de mettre en perspective des archives avec l’actualité du jour.

LA PRÉSERVATION DURABLE DES ARCHIVES NUMÉRIQUES

Le débat sur la préservation pérenne des archives audiovisuelles ne s’est pas clos avec le numérique. Les supports des fichiers numériques – qu’il s’agisse de disques (DVD) ou des bandes informatiques (DTF, LTO) – sont tout aussi vulnérables et mortels que les supports des enregistrements analogiques.

On a quasiment abandonné la recherche d’un format pérenne, immuable, stable. Seuls les archivistes du film considèrent encore aujourd’hui que la pellicule est, dans l’état actuel des solutions numériques disponibles,
le support le plus fiable et le plus économique pour la préservation à long terme des films 35 mm. Pour une conservation à moyen terme, ce qui est le cas à l’Ina des éléments film en attente de traitement, des mesures de préservation préventive sont prises pour le stockage (température de 10 à 12° ; humidité relative de 30 à 35%).

Le débat s’est déplacé sur un autre terrain. La solution pour assurer la préservation durable des archives numériques, c’est d’organiser la migration périodique des fichiers numériques sur de nouveaux supports. En effet la meilleure façon de conserver les enregistrements numériques, c’est de les recopier, à intervalle de quelques années.

C’est ainsi qu’en 2007, 150.000 heures de programmes numériscrits sur des bandes DTF2 ont été transférées sur des bandes LTO3. Alors que le passage de l’analogique au numérique avait pris plusieurs années en mobilisant d’importantes ressources humaines et financières, la recopie de numérique à numérique s’est faite en cinq mois et a été intégralement automatisée. La migration a aussi été l’occasion de faire le choix d’un support informatique « ouvert », c’est-à-dire préconisé simultanément par plusieurs constructeurs et largement implanté sur le marché, offrant de ce fait des garanties d’exploitation durables.

**LA SAUVEGARDE DU PATRIMOINE AUDIOVISUEL MONDIAL**

La coopération internationale entre les grands centres d’archives de télévision s’est organisée dès 1977 lorsque fut créée la Fédération Internationale des Archives de Télévision (FIAT), une initiative conjointe de la BBC, de la RAI, de l’ARD et de l’Ina. La question de la sauvegarde des archives a toujours été au centre des discussions, et lorsqu’il s’est agi de la numérisation massive du patrimoine sonore et audiovisuel accumulé depuis les origines de la radio et de la télévision, les archives européennes, très présentes dans la FIAT, ont mis en commun leur expérience, notamment dans
le cadre du projet PrestoSpace, un projet financé par la Commission européenne (voir l’article de Daniel Teruggi dans ce numéro d’IPN).

Au début de la décennie, et sous l’impulsion d’Emmanuel Hoog, élu en 2002 à la présidence de la FIAT, de nombreuses archives du Sud ont rejoint la Fédération. La coopération concernerait désormais aussi les archives les plus défavorisées. L’on a en effet considéré que la sauvegarde du patrimoine audiovisuel mondial était une responsabilité collective, partagée par la communauté des archives, et par les États, soucieux de protéger l’identité des peuples et la diversité culturelle. Il faut dire que la fragilité naturelle des supports, les conditions climatiques, les guerres, le manque d’intérêt et de moyens, font que ces archives sont infiniment vulnérables.

C’est dans cet esprit que plusieurs initiatives ont été prises pour alerter l’opinion et les pouvoirs publics sur la nécessité et l’urgence de sauver les archives audiovisuelles

- En octobre 2004, la FIAT lançait un appel mondial pour la préservation des archives audiovisuelles. L’appel qui a été repris par les autres organisations professionnelles internationales qui s’intéressent aux archives audiovisuelles, a recueilli le soutien de plus de 10.000 signataires dans 110 pays


- En 2007, Emmanuel Hoog lançait avec la FIAT le projet Archives en péril, un site web conçu pour illustrer, images et vidéo à l’appui, la riche diversité des archives dans le monde et pour sensibiliser l’opinion.
« L’objectif, écrit Emmanuel Hoog, était de créer les conditions permettant aux professionnels afghans d’assurer eux-mêmes, en toute autonomie, la conservation, la numérisation, puis la diffusion de leurs archives. Depuis sept ans, plusieurs milliers d’heures ont été numérisées et sauvées de l’oubli. Il reste encore à faire connaître ces documents numérisés, que de nombreux afghans n’ont jamais vus. Parce qu’ils ont été les acteurs de ces images, et qu’ils en sont aujourd’hui les uniques détenteurs, ils doivent pouvoir, dès demain, en être les spectateurs. Plus qu’une évidence, c’est une exigence que d’accompagner le peuple afghan à préserver et transmettre la richesse de ce témoignage sur son passé : c’est non seulement contribuer à la constitution de son histoire et de son identité, mais c’est aussi et surtout lui permettre d’écrire son avenir ».

Dans le même temps, l’Ina montrait l’exemple et engageait plusieurs projets pilotes de coopération, avec l’Afghanistan, le Cambodge, Madagascar, Timor Leste. On évoquera ici le plus abouti de ces projets, celui qui a pour cadre l’Afghanistan.


Deux missions de dix jours chacune, par an, à Kaboul, depuis juillet 2002 ont permis de procéder à l’expertise des fonds, d’apporter des équipements – lecteurs-enregistreurs vidéo, télécinéma, ordinateurs –, d’installer des chaînes de transfert et de numérisation, de former les techniciens à la maintenance et à l’exploitation des équipements, ainsi qu’à la gestion documentaire des enregistrements.

Les résultats (encadré):

- Radio Télévision afghane:
• Télévision : 25% des 15.000 heures à sauvegarder sont numérisés
• Radio : 25% des 30.000 heures à sauvegarder sont numérisés
• Afghan Films
  • 50% des 500 heures de film sont numérisés
• Ariana Films
  • 50% des 1.200 heures de film sont numérisés et disponibles grâce à un accord commercial sur inamediapro.com

UN PLAN POUR SAUVER LES ARCHIVES AFRICAINES

Aujourd’hui, les situations d’urgence sont d’abord en Afrique. C’est là que les archives sont les plus vulnérables, c’est là que la mobilisation des pouvoirs publics nationaux est la plus modeste, c’est là que les ressources sont notoirement insuffisantes.

On estime à 150.000 heures de télévision et 200.000 heures de radio les archives à sauvegarder dans les pays d’Afrique noire francophone, alors que le rythme actuel de sauvegarde n’est au mieux que de quelques milliers d’heures par an.

Et pourtant les archives de ces pays bénéficient depuis 2004 du soutien de l’Organisation internationale de la francophonie (OIF), du Conseil international des radios et des télévisions de l’espace francophone (CIR-TEF), et de la FIAT: participation régulière des responsables d’archives
africains aux conférences de la FIAT ainsi qu’aux sessions de formation du CIRTEF (SEFOR), déploiement dans dix sept pays d’un système de numérisation, documentation et de stockage des archives (AIME), donation de magnétoscopes ¾ pouce pour sauvegarder les cassettes Umatic qui constituent l’essentiel des fonds d’archives analogiques, etc.

Ces pays bénéficient en outre du soutien du ministère français des Affaires étrangères, à travers son Plan Images Archives. L’Ina, pour sa part, a initié un projet de coopération avec les archives malgaches et noué de nombreux accords avec les organismes de radio et de télévision africains pour l’organisation d’actions de formation.

Au total, il existe un grand nombre d’initiatives qui gagneraient à être mieux coordonnées. L’Ina espère pouvoir jouer un rôle décisif dans les années qui viennent afin de mobiliser les partenaires concernés, trouver les bonnes synergies, rassembler les ressources, et lancer, avant que la mémoire de l’Afrique ne se soit désintégrée faute de soins, un schéma directeur de sauvegarde et de numérisation du patrimoine audiovisuel africain.

Il s’agira de migrer les archives analogiques dans un format numérique, d’assurer la pérennité des enregistrements, de les rendre accessibles, de faciliter les échanges de programmes, et de créer les conditions de leur mise en valeur pour des usages professionnels, éducatifs, culturels et commerciaux.
CONCLUSION

Il existe un cercle vertueux qui unit la numérisation, l’accès, la sauvegarde : plus le volume des documents numérisés s’accroît, plus on peut diversifier les services d’accès, montrer les archives, générer de nouveaux usages, attirer de nouveaux utilisateurs, être « visible », intéresser les pouvoirs publics, trouver des ressources pour développer plus avant la sauvegarde et la numérisation. Il faut commencer le travail de numérisation sans attendre d’avoir réuni tous les moyens nécessaires à la réalisation d’un PSN « idéal », il faut montrer les trésors que recèlent les archives ; c’est un moteur efficace pour mobiliser, convaincre, avancer.

Le numérique modifie le statut de l’archive : le concept « d’original », inscrit dans la culture de l’archiviste des années analogiques, n’est plus aussi fondamental. Les copies numériques ne coûtent pas cher à réaliser ; elles n’ont plus pour effet de produire une image dégradée. L’authenticité du document original demeure, mais le numérique offre des outils de segmentation et de réagencement des contenus qui permettent de servir tous les usages. Le numérique confère à l’archive de nouvelles qualités : l’accessibilité, l’ubiquité, le partage. Il ouvre aussi de nouveaux questionnements : la protection des droits de propriété intellectuelle et la lutte contre le piratage ; la mise au point de nouveaux outils pour l’indexation et la recherche.

Enfin, et de notre point de vue d’archiviste ce n’est pas le moindre, le numérique a profondément modifié les métiers de l’archive. En particulier les documentalistes ont vu leurs pratiques changer. Les tâches traditionnelles que sont le catalogage, l’indexation, la recherche documentaire sont désormais largement réalisées par d’autres intervenants – le producteur en amont qui livrera l’archive avec ses données de description, l’utilisateur qui peut accéder en ligne aux bases de données et à la consultation des vidéos. Ces tâches se sont estompées pour laisser place à d’autres, centrées sur l’organisation du fonds, la valorisation des contenus. Des outils
nouveaux, on l’a vu, facilitent cette mission, et créent les conditions d’une connaissance fine, enfin maîtrisée des fonds, qui fait du documentaliste de l’ère « numérique » un spécialiste des contenus, un passeur. Dans ce monde numérique qui déborde d’images, ce rôle d’intermédiation est devenu essentiel.
Claude Mussou has a Masters degree in American literature and has graduated in bilingual journalism. She has been trained as a documentalist and working at Ina since 1987, coordinating web archiving activity since 2008.
SO WENT THE headlines of an article, published in The Telegraph on May 4th 2010, that reported on the release of IDC annual Survey of the world’s digital output. And it went on « …the current size of the world’s digital content is equivalent to all the information that could be stored on 75bn Apple iPads, or the amount that would be generated by everyone in the world posting messages on the microblogging site Twitter constantly for a century. »

Evidently, the digital universe refers to content available on the world wide web, 70 percent of which is, according to the survey, generated and published by individuals. Even more so, as numerous traditional publishing activities have been shifting on line, the sheer notions of author, publisher and public space have also been redefined.
SCOPE CRITERIA FOR WEB ARCHIVES: VARIOUS APPROACHES

In that context, and according to their remit or the scope of former collections, heritage or research institutions have developed, strategies, technologies and methods to collect and preserve online content.

The Internet Archive foundation which pioneered web archiving as early as 1996 has made a commitment to regularly harvest, preserve the whole world wide web and make it available to the public by means of the Wayback Machine (http://www.archive.org/web/web.php).

Many institutions, mostly National Libraries or Archives, collect web sites within the Legal Deposit framework and are mandated to harvest and preserve top level domain web sites so as to provide snapshots of their national web.

Within that framework, the nationality criteria prevails; however, mapping national frontiers on the web is not easy. It is indeed possible to have an agreement with national registration offices (AFNIC in France) that may provide on a regular basis an updated national domain list of relevant web sites, yet for instance, the .fr domain collected by BnF (1.600 0000 sites) represents approximately only 30 percent of the French web.

The approach might also be thematic or selective, either for heritage or research purposes, with institutions focusing on collections in relation to specific events, themes or activities (i.e. national elections, Olympics games, crisis and conflicts, environmental policies, natural disasters, public and government policies).

Defining a scope for collections is not an easy task when applied to the world wide web. The web is by all means not « frozen », it undergoes continuous transformation, with web sites appearing while others disappear. As highlighted above, it is large and deep, but also ephemeral, volatile, dynamic and interlinked. So much, that the selection process is a perpetual movement, relying on automatic tools for reporting and analyzing outgoing links, but also, and rather unexpectedly, highly dependant on human
surveillance. Indeed, librarians and archivists in most institutions are very much involved in nominating new sites and pointing to those, which are prone to disappear.

**THE FRENCH WEB ARCHIVES: A TWOFOLD BIAS**

In France, Ina, by law (IDADVSI, August 2006), shares responsibility with BnF, the French National Library, for the legal deposit of French web sites. In coherence with its core activities, Ina’s scope encompasses all web sites that somehow relate to the broadcast industry, broadcasters’ web sites evidently, but also blogs or fan web sites that focus on radio and TV programs or personalities. The BnF handles the rest of the French web, harvesting “snapshots” of the French web on a yearly basis.

As the web becomes a favorite platform for distributing videos and developing new formats (web documentaries, web series), collecting and preserving broadcast related web sites is of utmost importance for Ina so as to ensure continuity and coherence for its collections.

**INA WEB ARCHIVE SCOPE**

Definition of Ina’s initial scope was outsourced to Linkfluence, a French high tech company which performed the initial crawl and provided link and semantic analysis for the starting point of Ina’s web archive in February 2009. The harvest was not as plentiful as expected, notably because the seed sites did not outlink very much, due to the competitive dimension of a scope focusing on broadcasts.

Thus, the initially selected sites amounted to 3,600 when the archival process started. Rather narrow in scope at the beginning, the archive, 16 months after it started building up, points to over 7,000 relevant web sites
after a thorough and regular surveillance work is achieved by documentalists.

In a selective focus and rather restricted scope provides opportunities to go beyond the « snapshot policy » and perform very frequent and scalable crawls so as to attempt to replicate the dynamic dimension of the selected sites, “sticking” as much as possible to their editorial reactivity. Indeed, a specific and granular approach has been devised so that it is possible, according to a one crawler / one site policy to instruct each crawler on where to go, how often and how deep to capture each one of the site according to its size and update frequency. Home pages may be harvested as often as every two hours, and crawls may go as deep as 5 clicks away from the home page.

Naturally, the size and bulk of the archive is ever growing amounting at the beginning of June 2010 to 3,211,949,867 files/ URLs which equal 323,41 uncompressed TB. Naturally, audio and video files account for a great part in the archive (Audio 22,609,100 URLs / 122.79 TB and Video 11,784,198 URLs / 132.56 TB). The storage format (DAFF) is adapted to the crawl frequency and handles redundancy in files so that when content has previously been collected, only its metadata (location and time) is stored. The efficiency in storage is very crucial as 323, 41 TB equals 23 TB once deduplication and compression has been performed.

Many archives have started building a digital memory and are being confronted to core issues, notably linked to storage, preservation … and selection. The “archival attitude” has shifted from a tradition of comprehensive collections to the awareness that this is not achievable in the “digital universe”. Tolerate gaps, by pass traps and collect better rather than comprehensively, these are the new trends in building a web archive.
Irma van Kampen has a Bachelor’s degree in Business Administration, Master’s degree in Economics. She started working as a documentalist at the video archive of NOB, that is now Netherlands Institute for Sound and Vision, in April 1993. In 2005 she started as a copywriter, writing for the general public. These texts are published in the multimedia museum Media Experience, which is part of the Netherlands Institute for Sound and Vision.

Lisette Graswinckel has a Bachelor’s degree programme in Translation, English into Dutch, Bachelor’s degree and Master’s degree in Arts & Sciences. She has worked for the Netherlands Institute for Sound and Vision since 2004. In 2007, she became a senior documentalist. Since 2005, she is also a member of the research team for the Media Experience, a multimedia museum which is part of the Netherlands Institute for Sound and Vision.
DURING THE YEARS, the retention policy of the Netherlands Institute for Sound and Vision has changed alongside with practical and ideological changes in the field of audiovisual archiving. From a situation where we were thrilled with any tape or film that found its way to our desk, we are now in the happy situation to have all, or at least a lot of, the content ‘delivered to our doorstep’ by digital express. The present situation is that the amount of content we are receiving every day is larger than it ever was before.

This has led to a different kind of “selection policy”. Now, we can first harvest the programme, and then decide (or “select”) on what level of cataloguing – from basic to detailed – a certain episode should be archived. Our most important criterion in deciding this level is how to make the material available to as many catalogue users as possible. And where these used to be mostly media professionals, they are now as varied as
“the general public”, high school students and academics.

In this paper we will chronologically describe the changes in our selection and retention policy of both radio and television programmes and the practical and ideological developments that influenced these changes. As you will see, we used to harvest and catalogue radio and television programmes in different ways, but have now integrated the workflows as much as possible.

**THE PAST**

The Netherlands Institute for Sound and Vision is an audiovisual institute and archive, collecting film, video, audio, music, and photographs and artefacts from our broadcast history. Although it has always combined the particular roles of a company archive as well as a cultural historical archive, these last decades the emphasis has shifted greatly to the institute’s public function.

Until the start of the 1990s, when the television archive drew up a first official selection and retention policy, we were primarily a company archive. In those early days we found ourselves at the end of the production chain. We had no retention or selection policy, but mostly waited for the programme makers to deliver their programmes. Sometimes we were lucky and they brought us complete programmes, but sometimes they brought us only shots or items. All of these were archived and catalogued on a detailed shot level. However, looking at it with modern eyes, a lot of straightforward content information – let alone context information – is missing. A detailed catalogue description from those days is chiefly a sequence of time codes, followed by a short impression of what you see.

This passive attitude towards acquisition led to large gaps in our collection. This resulted in a growing discontent which, combined with our growing cultural awareness and the development of new technologies,
gave birth to our first official selection policy for our television archive in the early 1990’s. This completely changed our working process.

First of all, we no longer had to wait for material to arrive after broadcasting, but we started an active acquisition before broadcasting. Based on a list of criteria, we decided how many episodes of a certain series we wanted to add to our collection. Some of these criteria reflected our ambitions as an archive, other criteria referred to the content of the programmes, while still others were added to make sure we also collected the rare, the popular and the award winning programmes. In practice this meant that for some series, such as game shows, one episode per season was sufficient. From other series, for instance documentaries, every single episode was added to the collection. Second of all, we started to tape the programmes ourselves. This increased our independence from the broadcasting companies and ensured that we obtained the programmes we wished to have.

The next development in our selection and retention policy took place at our radio archive at the start of the twenty-first century. Due to cheaper harvesting methods and storage costs, we could retain more broadcasts (though still not all). At the same time we did not have enough personnel to catalogue every episode of every radio programme we selected in detail. This generated a new method of cataloguing, where not every episode was catalogued in detail, but only a certain percentage of every radio genre. This meant that, for example, 80 percent of the documentaries we had selected would be catalogued in detail, while only 30 percent of the music programmes, et cetera. This way, the archivists made the choice, based on all the metadata available to them, which episodes would be catalogued with a detailed summary and index words, and which episodes would not. Still, the episodes that would not be singled out (or ‘selected’) by the archivists would still be archived with all the basic information, such as title, publication date, genre, archive number, et cetera.
THE PRESENT

As we will see, the new idea of using different cataloguing levels mentioned above, has been at the root of our present working method with both radio and television programmes. Since 2006 (television) and 2008 (radio), we receive and store all or nearly all public broadcasts digitally at our archive. This means that the amount of content we are receiving every day is larger than it ever was before. The good news is that we can retain more content, and thus have to make fewer painful choices in what to keep and what not. The bad news is that we simply do not have the manpower to catalogue every item in detail, as before.

So, we extended the idea of cataloguing on different levels to the new digital situation, which meant, in a way, a different kind of “selection policy”. In the past, we used to make educated guesses about whether a certain programme or film would be interesting enough to add to our collection. This all happened before the programme was broadcast. Now, we can first harvest the programme, and then decide, or ‘select’, on what level of cataloguing – from basic to detailed – a certain episode should be archived. These levels are:

1. cataloguing the basic metadata with minimal effort;

2. cataloguing with more detail and more effort (with a smarter use of metadata and key frames);

3. cataloguing in detail with maximum effort (by listening to or viewing the complete programme).

The overarching principle in deciding this level is how to make the material useful to as many catalogue users as possible. Concretely, these criteria are the reuse value and the cultural historical value of the material.
In practice, the quality of external metadata also plays a role.

On level 1 we mostly find quizzes, game shows, compilations, gymnastic programmes, animations, and language courses. These programmes will not be edited by an archivist, which means they will only carry basic metadata and the metadata which the broadcasting companies have added before broadcasting. On level b) we find a variety of genres, such as soap operas, medical programmes, drama and talk shows. These programmes are catalogued by archivists, however not by viewing the complete programme. Instead, the archivist uses external metadata, which is checked against the content by using key frames or another way of spot-checking. This way, cataloguing takes up less time. However, the added information will contain less detail. On level c) the programmes are viewed or listened to from beginning to end, and a detailed summary and indexation are made. This is actually the old-fashioned way of cataloguing. On this level we mostly find documentaries, cultural programmes, serious talk shows, special broadcasts, et cetera.

There is another advantage to the digital archive. Because we used to select the programmes before they were broadcast, there was always the risk that we missed a certain interesting episode or programme. In the digital era we can work with hindsight, basing our decision how to catalogue a certain episode or series on the actual content. So, when we hear something remarkable on the news, for instance that one of the presenters of the Dutch television show Try before you die has streaked at Wimbledon, we can adjust the level of cataloguing and give this episode extra attention – because this is an episode people will remember.

THE NEAR AND FAR FUTURES

In 2006, the Netherlands Institute for Sound and Vision moved to a colourful new building. It not only houses our offices and collection, but
also the Media Experience, a public interactive media museum. Here, the
general public can visit all sorts of exhibitions. One of these is the Media
Lounge, where visitors can browse our digital collection. This meant that
a new group of people, as diverse as “the general public”, would be browsing our collection and read and use our catalogue entries. We realised that
our professional catalogue would not be appropriate, since the general public needs other information, presented in a different style. This is why we started to write specific texts for the general public as a new target group.

Moreover, the general public is interested in other programmes than the media professionals are. Where the latter often have news items at the top of their lists, the general public tends to be keener on shows and quizzes and drama. These are the programmes they remember from their childhood and would like to see again. Of course, there are more target groups that might be interested in our collections. We have already started to give some of these groups, such as high school students and academics, special access to our catalogue, often by way of special websites, sub catalogues, viewing facilities and tailored catalogue texts.

Since it would be inefficient to provide separate catalogue texts for every target group, we have started to develop the idea of a ‘core description’. This is a basic catalogue text which can easily be adapted to different purposes. We call this method “view once, write for many”. The archivist will have to view or listen to the material only once, after which he can create the core summary. This text can then be adapted to suit the needs of a particular target group.

In general, this new focus on other target groups than the media professionals also influences the choices we make while selecting and cataloguing programmes on a particular level. For example, a quiz show might not contain many stock shots, but still be interesting as a whole to candidates’ family members, or to teachers, historians, sociologists, et cetera. The definition of ‘reuse value’ is thus changing, or rather broadening.
Looking at the far future, it is hard to tell whether we will need a new retention and selection policy. What we do know is that we will collect more and more content. We also know that technologies will change and that the storage of media files will always cost money. Will technology bring these costs down fast enough so that we can retain everything? All these factors are uncertain. But when we retain everything, we can also learn and analyse how and what audiovisual material is reused. This way, we can better understand the true value of a certain programme or episode.

**CASE STUDY: LINGO**

To illustrate the above we will now present a case study of one programme title: Lingo. This daily game show, combining elements of Bingo and Mastermind, has changed little since its first appearance on Dutch television in 1989. There are different words to guess, different candidates and presenters, but on the whole the programme stayed the same. Of course, the obvious question is: what to do with this kind of material? Should we retain every episode, or is it enough to keep some examples?

Due to our selection and retention policies of the past – or rather, the lack of such a policy in the beginning – we have not collected all episodes of Lingo. In fact, the oldest one we have is the 100th episode. Fortunately, we started to actively acquire material for our collection in the 1990s, which means that we have more examples of those years. Since 6 October 2006 – the day we started our digital archive – we have retained every episode. And although these recent episodes are catalogued at the most basic level, we can select every out of the ordinary one and catalogue it at a higher level.

Ideally, every radio and television programme is retained and catalogued at the highest possible level. Of course, this can only be done when storage and maintenance costs are low enough. Still, one cannot guess the future
reuse or cultural historical value of a certain episode or programme. Some obscure quiz candidate could once become prime minister, or marry into the royal family.

CONCLUSION

In short, an ever-growing number of broadcasts and the wish to serve more user groups made us decide to retain more – preferably all – material, but to catalogue it on different levels (i.e. not all in detail). Our increased focus on our role as a public institute changed our retention and cataloguing policy. This development is summarised in the title of our presentation: “from collection to catalogue”, meaning that where we used to build a collection through retention, we are now building a catalogue through not only retention, but also a new way of cataloguing (“view once, write for many”). In other words, we are making our collections accessible and attractive for a larger and more diverse audience.
Since 2009 Göran Konstenius works as the head of the Audiovisual Department in the National Library of Sweden. Between 2004-2008 he worked as the head of the Acquisition and Documentation department in the National Archive of Sound and Moving Images, SLBA. Göran Konstenius has an educational background in cultural and behavioural sciences.
12 PRESERVING THE BEST METADATA AND CONTENT

THE TASK OF the Audiovisual Media department at this National Library is to collect and preserve material from radio and TV, films, videos, multimedia, as well as music and other sound recordings. Today the department’s collections contain just over 7 million hours of media material, of which about 2.5 million hours are stored as digital files in the National Library’s digital archive, accessible online via the department’s catalogue, Swedish Media Database. In 2008 the department collected just over 300,000 hours of TV and radio material as files delivered via networks, and it has a migration facility for digitising older radio and TV material with a capacity of over 2,500 hours per day. The department’s archive is thus very large and fast-growing, even from an international perspective.

Sweden has had a legal deposit law for books since the 17th century, and for audiovisual material since 1979. Until the end of 2008, the responsibility for collecting printed matter lay with the National Library, while the
National Archive of Sound and Moving Images (Statens ljud- och bildarkiv, or SLBA, in Swedish) was responsible for audiovisual material. As of 2009 these activities have been merged, and SLBA is now a department of the National Library – the Audiovisual Media department.

The department consists of four units: the Acquisition and Documentation unit, the Technology unit, the IT Development unit, and the Research and Web unit.

The Acquisition and Documentation unit is responsible for collecting and documenting material covered by the legal deposit law. The unit also receives material donations from businesses, institutions, production companies and private collectors. An important task in recent years has been to develop digital acquisition systems, in particular for broadcast material.

The Technology unit is responsible for analogue material and for devising preservation plans for that material. The unit also produces research copies for the department’s researchers by digitising analogue material, under a system of digitisation on demand. Some 10 000 media files are produced every year on demand from researchers. The Technology unit is also charged with setting up systems for the large-scale transfer of analogue material.

The Research and Web unit is responsible for external activities and for exhibitions, lecture series and web presentations about the audiovisual collections. The unit has recently developed a system that allows researchers to search the department’s catalogues and order material via the web.

The IT Development unit builds the applications necessary for digital collection and transfer, and for making the collections accessible.

Under the legal deposit law, the department must collect the following:

- All radio and TV broadcasts originating in Sweden, all Swedish and foreign films and videos distributed in Sweden, all Swedish music and other sound recordings, and Swedish multimedia.
Of the 60 staff posts that the department has at its disposal, three and a half are involved with the collection and documentation of television material. A further three posts are filled by IT and technology staff who work with digital storage and transferring older analogue material. In other words, the share of staff involved in television activities is relatively small. In terms of the volume of material collected, however, it is the department’s biggest activity. Currently it collects the full range of broadcasts from 46 national TV channels and 39 local TV stations.

SELECTION

The selection of TV material is based on the entirety principle, which means that all material broadcast by TV companies is to be submitted to the National Library under the legal deposit law. This applies to both national and local broadcasts. Since the law specifies that entire broadcasts must be submitted, selection has not been an issue for television archivists in the collection phase.

From 1979 until mid-2000 TV material was mainly delivered on videotape format, but over the last couple of years the delivery format has become almost entirely digital, sent in either via networks or as digital files on DVD.

Over the past five years the department has been actively involved in building a digital archive – from collection to enabling public access –, and these efforts have been centered around TV material, both in terms of abandoning the collection of analogue material in favour of digital collection, and in terms of transferring older TV material to a digital format.

The government has granted 20 million crowns – roughly two million Euros – towards the building of effective systems for transferring the public service material in its entirety. No selection was made in this context.
either, in keeping with the department’s task of acquiring and preserving everything. Using transfer systems developed by our own technicians, the department currently transfers about 2 500 hours of broadcast material per day. About 800 000 hours have so far been transferred using these systems. There are few institutions in the world, I think, that manage to maintain this kind of speed in their transfer work. Naturally, the quality of transferred material cannot be compared to broadcast quality, but then the department is a research archive and not a production archive.

QUALITY

Because the legislation does not contain any requirement for a specific technical standard, the technical quality of the delivered material is variable but generally low. The analogue collections are made up mainly of VHS recordings and the digitally collected material of mpeg1 and mpeg4, equivalent to – or slightly higher than – VHS quality. However, for the public service material there is a special cooperation agreement with Swedish Television under which the National Library acquires the company’s broadcasts of Swedish productions at a high quality level. In a cooperation project with Swedish Television, about 18 000 hours of this high-quality material has been digitised.

In this context, however, it is important to point out once again that the task of the Audiovisual Media department is not to serve as a national archive for Swedish TV material but as a research archive able to provide researchers with an overall view of the total supply of television material directed at or broadcast from Sweden. It should also be reiterated that this material is very extensive – today there are more than 45 national TV channels and nearly 40 regional ones whose broadcasts are being collected in their entirety. The National Library collections provide an overall picture, while the “national copies” of Swedish productions are held in the
various TV companies’ archives or in other production company archives.

The law does not prescribe any specific quality requirements for meta-
data either. Acquired metadata are updated programme schedules deliv-
ered digitally direct to the department’s catalogue. Some information is
added about the content of current affairs programmes and news. At pres-
ent this metadata completion work requires one and a half full-time posts.
Due to the vast extent of the material there are no resources within the
National Library to catalogue the material.

Being a research archive for television, the department’s archive is built
along modern lines. The digital archive is set up to allow the digitally
collected files and the digitally acquired or manually produced metadata
to be cross-referenced in a system known as the Swedish Media Data-
base. For TV material, this involves the collection of one-hour files of
TV broadcasts that are matched against programme schedules. When a
researcher searches the database and gets a hit in a programme schedule,
the system can play the file or files directly from the schedule item. How-
ever, for copyright reasons this system may only be used on the National
Library premises – which is one of our biggest concerns. Technically, our
2.5 million media files can be viewed on any computer in the world, but
copyright prohibits us from doing so.

AUTHENTICITY & THE ARCHIVIST’S CHANGING ROLE

In terms of content, the collected material may be regarded as having a
high authenticity value as it is collected in its entirety and includes con-
textual and other features such as hosts, presentations and signings off,
advertisements etc. In purely technical terms, however, the collected ma-
terial does not reflect the original broadcast quality. Furthermore, as the
material is not catalogued, metadata is not checked to a satisfactory extent.
In contrast with the acquisition of other audiovisual material by the department, the process of collecting TV material is highly automated— which we perceive as the future of media archiving. Files of metadata and TV broadcasts are delivered automatically and cross-referenced in the Swedish Media Database. Considering the enormous supply of material, I cannot see any possibilities for collecting and preserving it in a satisfactory way without further automating processes, in particular documentation. As everyone knows, systems for the automatic indexing of digitised sound and video material already exist, but the technology remains very expensive and has not been fully developed—particularly where systems adapted to the Swedish language are concerned.

Over the last few years, focus has increasingly shifted to developments on the Web. There is today no law requiring TV broadcasts on the web—“webcasts” or web-TV—to be submitted to the National Library. However, more and more TV material is being webcasted. Some of this is material is from the established TV companies, but a large proportion comes from other players such as associations, companies and more or less private players. During spring 2009 the Ministry of Education and Research has begun an official inquiry into the possibility of a legal deposit law that includes the Web. Myself and other National Library officials are members of a reference group reporting to this inquiry, which will produce a proposal for a bill to be circulated for comment this autumn. The background is of course that more and more material is being distributed over the Web—arguably, the future of television lies on the Web—and this material will be lost to researchers if it is not acquired.

Hence, when it comes to television and the Web, the archivist is faced with a tricky and very crucial issue: how do you distinguish between TV and other moving images? We are moving from a clearly defined TV medium with television sets and airwave broadcasts towards various types of video clips and direct-streamed files over global networks. Defining and
collecting all Swedish TV material on the Web is a huge challenge. In our case, as I have mentioned earlier, all video and TV material not distributed over the Web has been collected based on current legislation. On the Web, distinctions between types of media often disappear. Is it TV or video that is being distributed? Should the leading TV companies be viewed increasingly as content producers – not just of what we today regard as TV, but of video, text and interactive applications of different kinds?

In the strictest sense, Swedish legislation defines only live broadcasts as TV. Television companies do broadcast via the Web in this way, but in principle live broadcasts of company information gatherings, political speeches or debates – as well as images from cameras placed at various events – are all TV according to this definition. And such webcasts are becoming increasingly common. For example, the traffic outside this building can be monitored via a website that shows live images of the traffic situation in Stockholm. Is this television?

Even if new legislation were to limit the task of archivists to acquiring video and TV material distributed via traditional TV companies or other larger media companies, as it is likely to, the entirety principle will be challenged by the character of the Web. Collecting files distributed over the Web by TV companies would not provide a full picture of the new TV medium as it is being offered to the general public via the Web, mobile telephony or other IT networks. The contextual elements which used to be acquired and which added authenticity to various media would disappear if only the video files themselves were collected. Alternatively, entire websites would have to be collected, with all the updates made to them and all the functionality built into them. Currently, the technical conditions do not exist either for companies to deliver live websites or for the National Library to collect or acquire them and make them accessible.
From the archivist’s perspective, then, activities are moving from the acquisition of fairly easily defined material that can be collected in its entirety, towards a new media landscape where it is increasingly difficult to define what constitutes a TV broadcast, and where the entirety principle will have to be abandoned and a large part of the context lost. The selection of material irrespective of whether it is broadcast live via networks or distributed as video clips, will thus require a whole new set of competencies for the archivist of the future. Even if the guidelines for what is to be collected are laid down in the new e-deposit law, usage and interpretation of the new law will have to be established, and the archivist will have a very important role in this process. The archivist’s professional role will become a much more proactive one, and will require considerable knowledge about developments on the Web as well as on the Internet – and how these relate to the overall principles for what is to be collected, as well as knowledge about researchers needs.

A final remark is that the development of systems for acquiring Web material – irrespective of whether it is delivered by the producer or actively downloaded from the Web or other networks – will definitively need to be done collaboratively between archivists, researchers and IT developers.
PRESERVING THE BEST METADATA AND CONTENT
Since 2007, Eric has been the Chief of Logistics at TFO in Canada where he is responsible for the implementation of the organizational workflow around the HD tapeless environment. Prior to that, he worked for Canal + Group on the standardization and optimization of the productions’ process in Europe and was the Director of Operations for Canal + Group USA, responsible for the moving of the TV channel Eurochannel from Sao Paulo to Miami. Eric has a bachelor degree in Broadcast Television from Université de Provence.
TFO IS a public French language educational television broadcaster in the province of Ontario. It is the only French-language television channel in Canada whose operations are based entirely outside of Quebec. Launched in 1987 as “La Chaîne” and originally a part of the TVOntario, “TFO” was renamed using our current name in 1994. On April 1st 2007, we became a fully autonomic organization.

Our head office is in Toronto, with two regional offices in Ottawa and Sudbury. TFO has 150 employees and a yearly operational budget of 25 millions dollars. Our channel has 21 hour airtime every day, supplemented by 3 hours of nightly downloadable content for use by teachers in the classroom. TFO also has a strong presence on the new media platforms, with more than 10,000 videos amounting to 2,000 hours of content available thru the web, the podcast and the cellular.
TFO’s programming can be divided into 3 sections:

- In-house production: We produce 200 hours of fresh contents every year, ranging from live shows for kids, a daily live current affairs show, a teenager show and segments for 4 different groups’ ages (2 to 6, 7 to 12, 13 to 19 and adults).

- Co-production: Collaborating with external francophone production companies in both Ontario and other Canadian provinces, every year TFO co-produces around 100 hours of content, ranging from drama series, sitcom, documentaries, cartoons, as well as children series.

- Acquisitions: Representing more than 200 hours of programs per year, acquisitions complement our programming especially with our nightly rendezvous of art-house cinema.

SO LET’S GO BACK TO APRIL 1ST 2007

The government implements the TFO autonomy. For the past 20 years, we were living in the same environment with betas SP composites facilities for studios and assembling suites. A few years ago we added non-linear Avid edit suites. The total number of legacy tapes was round up at 55 000 beta SP. I know it was not a lot for some of you, but for us, it was a large volume of assets to manage.

On the database side, it was worse! We used more than 40 different databases to manage our operations from scheduling, traffic, tape library, copyright, studios run-down… etc… We were using a variety of software ranging from Oracle to several version of FileMakerPro as well as excel files! Some of these databases were connected, while most of them were
not. Metadata were often all over the place, sometime redundant in 4 or 5 places and not always accurate. This was one of the major problems to handle. How do we find the accurate data? Which one is relevant or not? While with our former owner, TFO had access of an I.T. team of more than 30 peoples as well as many skilled programmers; the new independent TFO does not have that luxury since we have only 6 staffs dedicated to both the I.T. and Broadcast department.

The first challenge was the timeframe. The government implemented the autonomy on April 1st 2007. We had to be out of the building in January 2008. The new independent TFO had only tapes and the people. It left us 9 months to find a new location, to make the construction, to build the technical facilities, to buy the equipment and furniture, to do the testing, to design the new workflow, to train the people… etc… and we had accomplish all these tasks without interrupting the on-air broadcast.

**OUR PLAN**

The first thing we decided to do was “to put aside 20 years of work practices”. Do not try to reproduce exactly what you had been doing, but instead understand what your goals are and find new ways to achieve them. So many people are doing tasks mechanically without knowing why. I am always telling a short story to people about that:

One day, a friend of mine asked me why I was cutting the edges of a roast beef before I roasted it. I told him:”I don’t know! My mom showed it to me that way.” So I called my mom to ask her. She told me:”I don’t know! Your grand-mother showed it to me that way.” So I called my grand-mother to ask her:”Why do you cut the edge of the beef before roasting it?” She answered:”it is because my oven is too small”.

So try to understand why you do something. Since the workflows are always changing, your tasks have to change too. It is difficult for people
to adapt to the new ways. We like to stay in our comfort zone because we know it. We don’t like to go outside our boundaries because it may put us at risk. This is the role of the company to help people adopt their new tasks and the new workflow. At TFO, we were lucky since we were reorganizing the channel without taking anything from our old parent channel, we had no choice but to rebuild the workflow from scratch but people lost all their references.

The second thing was to start with one centralized database to manage all of TFO’s operational needs. As I told you earlier we had a very small I.T. and Broadcast department and the way it was structure before did not make any sense for us since we were able to start from scratch. Also, one database will help us to organize the HD tapeless workflow around it. We thought of the structure of the centralized database first then we built the technical infrastructure around it.
The last thing was to be able to manage automatically the distribution of the contents on several platforms. We wanted to put the web, podcast and cellular at the same level as the on-air broadcast; especially since it is getting almost impossible to have a teenager sitting in front of the television to watch a show.

Our head office facilities are composed of two studios and a control room; three Final Cut Pro editing suites; four EFP Panasonic P2 cameras; one audio mix suite; an ingest room with two workstations; one Graphics and one Character Generator room.

After a request for proposal for the centralized database, we decided to choose “Louise” from ProConsultant. With this product, we were able to replace most of our former database; do the scheduling for on-air, VOD, web, podcast and cellular; manage the copyright and the rights for acquisition, co-production and in-house production; manage the tape library and also manage and automate the HD tapeless workflow.

Left is a very basic workflow of the plant. The brain, it is how I call it is the asset management software, Louise. It is the entry point of the tapeless environment for all the end users (on the top left). In the middle, we have what I call the arms of the system, Front Porch. It moves the video files from the archive (on the left) to the production environment (on the right). Then we have the metadata workflow from the end users to the asset management software, to the edit suites, studios, and broadcast platform (cellular, podcast, web and on-air). Instead of using many different databases on different software, our centralized system meets all our operational needs. Our end users, instead of using different software to perform different tasks, now only have to be familiar with one software to perform most of their tasks.
Table 1: Accuracy of Metadata

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups of users use siloed, partially redundant databases</td>
<td>All users use one database</td>
</tr>
<tr>
<td>No consistency in data/data types</td>
<td>Standardized data structure</td>
</tr>
<tr>
<td>Mistakes affect few users</td>
<td>Mistakes affect all users</td>
</tr>
<tr>
<td>To use data, users would print it from one system and retype it into</td>
<td>Data is readily usable</td>
</tr>
<tr>
<td>another one</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Moving to a Tapeless Environment

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can keep the tapes of my shoot under my desk</td>
<td>Video files are intangible and safely</td>
</tr>
<tr>
<td></td>
<td>stored on a server</td>
</tr>
<tr>
<td>I am an “artist”, I don’t need to know anything about computers</td>
<td>I cannot go on a shoot or prepare for</td>
</tr>
<tr>
<td></td>
<td>post-production without a computer</td>
</tr>
<tr>
<td>I go to several places to prepare for post-production, ingest for the</td>
<td>I can do everything from my desk</td>
</tr>
<tr>
<td>web, broadcast the show</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Operational Grey Areas

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>The IT and Broadcast departments were 2 separate entities</td>
<td>IT and Broadcast staff are integrated</td>
</tr>
<tr>
<td></td>
<td>into one department</td>
</tr>
<tr>
<td>We had people managing the content and people managing the container</td>
<td>A journalist can publish his video on</td>
</tr>
<tr>
<td></td>
<td>the web without the help of a technician</td>
</tr>
<tr>
<td>We had a stock shot department who was reviewing all the footage</td>
<td>A journalist can select the stock shots</td>
</tr>
<tr>
<td></td>
<td>from his shoot to share with everyone</td>
</tr>
</tbody>
</table>
We had two golden rules:

- “The information is entered by the person who has the information”. It could be a difficult one. Not everybody is willing to go into the system and enter the information needed by others. The person can enter the information she needs but why should she enter the information needed by others? To implement this rule we had the support of all the directors of the department.

- “The information is entered only once and shared”. This is where you can get the efficiencies. But this rule cannot go without the first one.

Now, I will show you some examples of the impacts and the changes that were made with the implementation of the fully tapeless environment. On the few next slides, I divided the tables in 2 sections. The left column is for facts that happen before the implementation and the right column is for facts that happen after the implementation.

**TABLE 1: ACCURACY OF METADATA**

It was a challenge to identify which metadata to choose from one to the other since we have a lot of redundant information in several databases. The metadata were not always the same and it was difficult to have a constant quality of information.

With only one database, we are now more efficient in sharing the information. People have to be more responsible for the data entries since everybody is affected by any mistake in the system.
This is a major change in our environment. Going from a composite Beta SP plant to a fully HD tapeless environment. The biggest impact is the first one. Now we are dealing with intangible video files. We have to manage more than 20,000 new files every year. You must understand the workflow if you want to achieve your task. Since everything is managed by the asset management software, you can get trapped somewhere if you forget to do an operation and you cannot perform your work. Now, you need to be computer skilled at a good level. It also reinforces a team work spirit since everybody in a group is helping each other.

Since we were starting from scratch, we could design the organigram of the company the way we wanted. The major decision that was made was to merge the I.T. and the Broadcast department together. Many technical equipments now fall into this grey area since it is I.T. equipment used to perform broadcast task. Many broadcasters still have problems deciding which workflow to adopt because there are 2 different departments involve in the vision and the decision to make.

We are still in the process of making changes and adjusting the workflow but I can say we are 100 percent tapeless and are very happy to have made that step.
AN EXAMPLE OF NEW WORKFLOW

A JOURNALIST HAS TO SHOOT A SEGMENT FOR OUR CURRENT AFFAIRS SHOW AND PUBLISHES IT ON THE WEB:

1. Create the project in the Asset Mgt Software (AMS) and export metadata onto SD Card
   - Journalist

2. Import metadata into camera
   - Camaraman

3. Shoot
   - Journalist

4. Bring P2 card to ingest room
   - Journalist

5. The AMS links the video files back to the project, creates the proxies and archives the rushes
   - Ingest Operator

6. Import P2 card into the AMS
   - Ingest Operator

7. Upon edit, the AMS archives product and creates the proxy
   - Editor

8. Import EDL into Final Cut Pro
   - Editor

9. Upon approval of the final edit, the AMS erases the video files from editing
   - Producer

10. Entering prompt: change, synapse... info
    - Journalist

Before, during or after...

- Determining the version needed (web, podcast...) for automated publishing
  - Producer

- Preparing the run down of the show
  - Production assistant

- If not a live, enter date and time of studio shoot so the AMS knows when to copy the video files
  - Production assistant
CHANGING ROLES IN A TAPELESS ENVIRONMENT
Jacqui Gupta has worked for the BBC (British Broadcasting Corporation) more than 25 years in various roles both in production, news and archives management. Her current role is Lead Technologist in the technology consultancy team of Future Media & Technology with her expertise of digital media asset management and metadata within production and archives. More recently, Jacqui has participated in several EU collaborative projects and currently working on the EU PrestoPrime project. She is a proactive member of the Media Management Commission of FIAT/IFTA and a member of the Seminar Programme Committee. Over the years, she has given several presentations at IBC, FIAT/IFTA, EBU and DAM Symposiums and also contributed to SMPTE standardisation activities.

Jacqui holds a BA degree in Spanish and European Studies and an M.Lib Masters degree in Librarianship and Information Science.
OBSERVATIONS

THIS SEMINAR COVERED a diverse range of issues, presented new and challenging perspectives and fuelled an ongoing inspirational debate. However, it is really difficult to derive really concrete conclusions from the presentations and activity sessions – more a cluster of observations and even more questions.

GENERAL

There are very different missions for broadcast and national archives and small and large institutional companies. This can cause some confusion when trying to align missions from each sector. In the broadcast archive, the main priorities for selection are repeat transmissions, re-use value and cultural heritage. In several presentations it was stressed that the value of
the archives have never been so high due to many services on many differ-
ent platforms, eg iplayer, open archive, iptv and mobile devices. Therefore,
in a rapidly changing broadcast environment it is important to develop
new retention policies. For a national archive, the cultural heritage aspect
will be the primary focus in developing and maintaining a selection policy.
Increasingly, the size of an organisation will be an important factor when
developing a strategy for an archival policy. Large companies may be in
a better financial funding state to invest in staffing and resources to sup-
port a sustainable archive through well developed selection policies. Un-
fortunately, smaller companies may not have the funding, or indeed, the
management support to maintain a professional archive with professional
librarians.

This led us to some serious questions on definitions. What is selection?
What is retention? What is everything, what is something? Again, no clear
answers.

Keep it? Select it? Retain it? Not selecting before meant destruction,
simply because content that was not kept in the archive usually deteriorat-
ed in a producer’s cupboard. However, not selecting now…it will still exist
somewhere and be accessible in some form eg web/online for more users.
Also, keeping everything can be problematic especially when dual running
both a legacy and an online archive. Keeping the best as much as possible
is the main purpose of a broadcast archive because it is impossible to keep
all the raw material. Therefore, selection and retention is paramount to
enable a living working archive.

**SETTING UP NEW SELECTION POLICIES**

There was a concensus that a digital environment makes you rethink your
corporate retention policy. Every archive should have a selection and re-
tention policy based on cultural and re-use value. Our ideas on selection
are based on defined concepts (programmes, products, collections) and
do not take into account the distributed linked character of digital data.
Automated retention technology within media asset management systems
has enabled us to develop and improve a more PROACTIVE archive
management policy of more complexity. Now we have outgrown corpo-
rate policies which were based on transmission and linear workflows. In
the new digital age, policies will mature as our experience increases.

We have a dilemma which raises the question – do we keep everything?
Digital working is very immediate. There is a need for prioritisation as
resources are heavy. Do we keep material safe as an archivist or is it based
on cost? Storage, compression technologies and new formats in the form
of HD and 3D will be expensive. Therefore, there is a compelling need
to re-examine what is really important to your organisation. It is better to
select 10 good programmes, inserts or clips than 20 bad items. Should we
keep all versions of the same clips? Some broadcasters are obliged to keep
a copy of all broadcast material including versions. However, how many
copies do we keep? Again, this will depend on the organisation and type
of archive.

WHAT TO KEEP?

Selection is all about access especially with increasing volumes of content
(media and metadata), the ever changing dynamics of production, dif-
ferent users and ways of using the Web. Access is enabled by new evolv-
ing technologies, advanced search engines, user generated metadata and
social tagging. In a digital production chain archiving is all over the place
not at the end of the process but all the way through the process. A big
issue is that we do not have a clue about the heritage value of content at
the beginning of the process. We cannot decide which information will be
important in the future. So what do we keep?
New trends are highlighting the question of Quality before Quantity? It is more valuable to keep quality images which will be re-used and represent a cultural moment. Bad shots will not be re-used or used simply because there is nothing better. The value of content to an organisation must be recognised in the selection and retention policy. Also, the metadata is key – it is more significant to keep 5 clips with good documentation than 10 without any at all.

A new paradigm was hotly debated. “Keep everything and find/select later!” Academic researchers believe that it is important to keep everything in a new data mining and aggregation society which need as many sources as possible. There is a new digital divide with the need to balance “social and cultural best practice and important and non-important culture. Data mining is becoming the social and cultural way to solve the dichotomy of everything or something. So, in terms of cultural heritage the criteria will change and we should keep some garbage! Conversely, only 10-20% of what is kept is searched and re-used.

Quite interesting comments on news environments – Journalists decide what to keep and Archivists decide what to throw away.

**STORAGE ISSUES**

Some key points were raised and discussed and again there were more questions than answers. There are new physical storage issues in a digital world – do we store on data tapes in robots or on shelves? Robots are costly and access time is short. Also, in order to access one clip out of 100 the whole tape has to be stored. For long term storage 20 percent could be stored on shelves. We must always have a back-up tape strategy which is twice as expensive.

Do we keep everything in broadcast quality? What about browse quality versions? It really costs too much to keep everything at the moment. So
is the cost of storage really decreasing? Moores Law dictates that storage costs will go down so we could keep more and not less. So HD storage will not be a problem in the future?. A few considerations to take into account – there will be mass storage systems which will require investment, licences, service contracts and 10 year change solutions.

Some questions will always remain. What will our storage policy be in ten years time? Will we still have tape stores underground in the vaults?

**METADATA**

Times have changed! Archiving is no longer at the end of the production chain due to new digital workflows and processes. Within the broadcast production environment production teams/programme makers are creating and entering metadata early in the production cycle and staff are inputting metadata early in the production process. This raises issues of metadata quality/assurance and more importantly the need for metadata quality criteria. The role of the metadata archivist is evolving to one of training and updating user manuals, validation using a set of rules and adapting and reinforcing those rules to different user groups. Again, there was a consensus amongst the delegates that the validation of metadata authenticity should remain in the hands of the archivist.

Digital workflows create more metadata scrubbing by archive specialists not less. What is metadata to what level? Not all relevant metadata is kept but not all good metadata is kept. So, in terms of sustainability, do we delete excess metadata? Or delete because of a lack of metadata? Sustainability of metadata in selection and retention management is another key challenge within archives. Metadata integration is still a cutting edge feature within a broadcast production archive infrastructure workflow.
New search tools can assist in the creation of metadata (eg speech to text, scene detection, content based image retrieval) but the technology is still not perfect and requires professional human interaction. Rich media navigation techniques are user driven and there will still be taxonomies and ontologies in the future. There will be more metadata in the future and manufacturers are striving to understand the intent of the user in the right context in order to develop new metadata and taxonomy tools. Emerging rich media navigation techniques and improved search techniques will increase findability. Metadata is about finding what you want!

"Use Power of the Crowd"? – user generated metadata (UGM) is an emerging alternative source of metadata which is growing in the web/online space. Interactivity, collaboration, sharing and social tagging is becoming more common within online user communities. This could be a source of additional metadata for both broadcast and national archives.

Should we preserve all legacy metadata? Do we archive the internet web pages? How do we archive? Some of the main issues are that the web is intangible with multiple producers and it is impossible to measure the volume. An example of a selection policy could be to archive by depth on a daily basis (two clicks on each item from a home page). Indexing is important for efficient search and retrieval and each site would need to be validated by a documentalist. Web archiving is proving to be a real challenge to archives.

SOME ANECDOTES

Digital workflows have created one of the most seismic shifts in corporate archiving philosophies.

The value of the archives are at their highest level in history with repurposing to multiple platforms
Researchers say keep it all! And the professional archivists need to select and cannot keep everything practically

Metadata is important but content is more important. Professional Librarians and Editors will always be needed

Jacqui Gupta, BBC
<table>
<thead>
<tr>
<th>PAGE</th>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1028856 – I VALET OCH KVALET PÅ FIKASTUNDEN</td>
</tr>
<tr>
<td></td>
<td>1960: American actress Martha Hyer having a hard time choosing at the press conference of her film “Herrin der Welt.” © SVT Bild</td>
</tr>
<tr>
<td>20</td>
<td>210724 – FRÖKEN TV 1950</td>
</tr>
<tr>
<td></td>
<td>1950: Miss Television 1950 © SVT Bild</td>
</tr>
<tr>
<td>21</td>
<td>1038717 – FIDEL CASTRO BESÖKER SOVJET 1963</td>
</tr>
<tr>
<td></td>
<td>1963: Watched by Mr Khruschev, the Russian Prime Minister, Cuba’s Premier Dr Fidel Castro takes a snapshot in Moscow. © SVT Bild</td>
</tr>
<tr>
<td>28</td>
<td>1020167 – HARRY SÖDERMAN I LABORATORIET</td>
</tr>
<tr>
<td></td>
<td>19??: Swedish criminalist, who is the inventor of a new vacuum-cleaner for criminalists laboratory work, in work with the new apparate. © SVT Bild</td>
</tr>
<tr>
<td>29</td>
<td>1072198 – ANTROPOLOGIFORSKNING</td>
</tr>
<tr>
<td></td>
<td>1963: Soviet Anthropologist Prof. M. Gerasimov at work. © SVT Bild</td>
</tr>
<tr>
<td>36</td>
<td>95449 – STORDATOR</td>
</tr>
<tr>
<td></td>
<td>1949: Huge calculator aid in atom research, Princeton Univ. © SVT Bild</td>
</tr>
<tr>
<td>37</td>
<td>82761 – VÄG PÅ VILLOVÄGAR</td>
</tr>
<tr>
<td></td>
<td>19??: Street torn up by landslide, San Francisco, USA. © SVT Bild</td>
</tr>
<tr>
<td>50</td>
<td>992596 – FOTOGRAF PÅ HÖG NIVÅ</td>
</tr>
<tr>
<td></td>
<td>1932: A photographer on one of New York’s skyscrapers. © SVT Bild</td>
</tr>
</tbody>
</table>
51 1133325 – VILNIUS
2007: Film archive with thousands of film reels from soviet times, Lithuanian Radio and Television, LRT. © Lars-Gunnar Bengtsson/SVT Bild

74 158459 – KONTROLL
1947: The control of disease is achieved through knowledge. © SVT Bild

75 726615 – MAGNETISK PLAST
1959: New plastic magnet introduced in Marietta, USA. © SVT Bild

128 59316 – UNDERVERKNANDET AV VAPENSTILLESTÅND
1954: Colonel Ta Quang, Vietminh defence minister, sign agreement with France to halt the Indo-China war. © SVT Bild

129 856094 – VIDEOBANDSPELARE
1962: 2" video tape machine at Swedish Television, Stockholm. Installed, when new, summer 1959. © SVT Bild

136 1196207 – KLIPPBORD
1984: Television archive assistant, Kurt Boman, marking scenes for copying at Swedish Television, Stockholm. © SVT Bild

137 1794879 – KOPIJETJÄNST
1984: VCR video tape machine operation at Swedish Television. © SVT Bild

144 1802090 – I ARKIVET

145 1802089 – I FILMARKIVET
1962: Swedish Television Archive, Stockholm. © Bo Arrhed/SVT Bild
162 **11463 – ARBETSSTUGAN I TORNEÅ**
1939: Unpacking and selection of 2nd world war civilian aid from sweden, Torneå, Finland. © SVT Historiska Bildarkivet

163 **176266 – FRUSEN PUBLIK**
1938: Two spectators sheltering from the rain in the centre court at White-man cup, Wimbledon, UK. ©: SVT Bild

174 **1894866 – MELODIFESTIVALEN 1973**

175 **658191 – ARKIVARBETE**
19??: Archivist at work. Caption missing © SVT Bild

198 **171187 – ARKIV FÖR MAGNETBAND DAFA**

199 **1793234 – FILMDESTUKTION**
198?: Archive assistant, Harry Hammarlund, destroys rejected film, Swedish Television Archive, Stockholm. © SVT Bild