Preservation of the web
Issues for audiovisual archives

FIAT/IFTA General Assembly Antalya, 13-18th of October 2002

Annemieke de Jong
Netherlands Institute for Sound and Vision
FIAT/IFTA Media Management Commission

CONTENTS

I The infinite library

In 1941, the Argentinian writer Jorge Luis Borges, wrote a story called “The Library of Babel”. In this story Borges imagines a universal library, containing all human knowledge.
Its contents are endless and eternal. This library universe is composed of an indefinite and infinite number of shelves and galleries, and contains all books ever written, in every language of the world. One can find anything here: writing in every genre, information for every purpose, guidance for every problem. All works of history, science and arts.
At first the Library appears a strange and beautiful dream: “When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. All men felt themselves to be the masters of an intact and secret treasure”, the story tells us. But very soon the dream turns into a nightmare: it appears that the Library has no perceivable order. There is no index. Somewhere in its endless shelves a catalogue sits, but no-one can find it. What’s worse, amidst numberless shelves of useless books, a series of false catalogues can be found, that mislead the user and are unable to provide real access to the collected treasures.
It is a disaster: in the Library of Babel all human knowledge, all cultural products of past, present and future are there for the taking, but no one can touch them. The Library, which contents represent mankind and its quest for wisdom and culture, becomes a symbol of chaos and ignorance.

Borges’ story can be easily used as a concept for cyberspace as we know it today: a place where all knowledge is collected and all boundaries have transcended. Text, audio, video, fiction, non- fiction: past and present genres and resources merged into one huge multimedia collection: the web as the true and infinite media archive. The lost catalogue represents the dark side of the web, and can be interpreted as a methaphor for the lack of structure, the missing order, the thousands and millions of scattered and unarranged documents and materials. The false catalogues in the story are the area’s on the web that are lost or incomplete. They refer to the obsolescence and the incompatibility of formats. They illustrate the gaps, the overload and the overlaps, the uncontrollable and random web collection.

II Digital heritage

In a way these two angles - the promises, the opportunities of the digital domain on the one hand and the threat of chaos on the other - are reflected in the resolution of the Council of the European Union, dated the 25th of June of this year. This resolution, no. 2002/C162/02, can be seen as a prelude to the Six Frame work programme, that will start next year. The Council firmly announces to work on a new plan for European cooperation in the field of preservation of our digital cultural heritage. The resolution reveals an open eye to the shadow of chaos and ignorance that cyberspace casts. There is an emphasis on the vast and ever changing methods of creating, storing and preserving records. The need to continually develop methods for broad access and long term preservation is stressed. European member states are urged to join forces and cooperate through frameworks and the relevant custodial organisations such as archives, libraries and museums, must be helped to extend their research and exchange of ‘digital’ skills and knowlegde.

The Council explicitely mentiones the collections of broadcast organisations as a part of the digital heritage. It notes that *‘the digital preservation extends the existing vast publicly held collections, and that a significant quantity of digital content is also in possesion of broadcasters, which should be taken into acount when analyzing the situation and planning long-term preservation measures*’.
This ‘significant quantity of digital content’ evidently refers to our radio, television and stills collections, parts of which are now either being digitized or digitally produced. Following the resolution of the Euopean Council this ‘quantity of digital content’ holds yet another category: websites. Many broadcasters been producing impressive amounts of this content since1995. The websites in some cases, have evolved into complete, interactive multimedia universes, containing all sorts of linked information associated with broadcasting, news, culture and entertainment. These new-media publications can be seen to develop into a company collection in itself, and should –according to the resolution- be evenly included in the preservation measures.

Today I would like to explore what is being done and what could be done, regarding the long-term preservation of webcontent, produced by broadcast organisations. In particular I would like to map out the position of the audiovisual archive here. Can we connect our mission, our collection policies and our archival standards to the domain of web preservation? Should we? Do our specific archive principles and values fit here in the first place? And if so, why?

III Who preserves the web?

Preservation of cultural heritage such as books, legal documents and other publications has traditionally been in the hands of national archives and libraries, supported by legal procedures and formal criteria. Audiovisual archives – be it broadcast archives or national archives- are seen as specialized archives that have specific responsibilities for the core of the audiovisual part of this heritage: the films, the video, the audio and the stills. Websites are multimedia documents that combine all types of content. Many websites are dynamic and have different functionalities. These mixed objects do not fit into the traditional categories. On the basis of existing policies it is not easy to decide where the primary responsibility for collecting and preserving these materials should lie.

National libraries generally approach the www from the angle of deposit legislation. Online electronic materials are treated as an extension of print publishing. Several libraries have developed strategies for actively selecting and preserving websites, on the basis of the concept of ‘publication’. The Pandora project of the National Library of Australia is an example. The National Archives of Australia, the Library of Congress and the British Library have extended policies for electronic record management, to include websites. In France the National Library and INA have joined forces in this area. Another type of national effort is the Finish EVA project, that collects all freely available published static HTML documents, with their inline materials, in the Finnish domain. In Sweden, the Kulturarw3 Heritage Project is capturing Swedish sites and all other webcontent on Sweden.

These preservation projects strive to develop techniques to capture content, functionalities and ‘look and feel’ of their numerous objects, in a way that fits their collection policies. They either take selective snapshots of the site as a whole, or try to manage its separate components. Storage is often done on stable carriers like cd-rom and dvd. Frequent updates in many cases have to be taken into account. The biggest issue is often the choice between an object-based or an event-based approach. To what extent should the individual client-server-transactions be saved? Or might it be wiser to hold on to the objects that constitute a site on a given moment?

In all of these actions, the centre of the collection policy is the idea of national production, constituting cultural heritage: the sites should be about the country and/or written by countrymen.

To the harvest robots of the Internet Archive, based in the United States, all webcontent is equal. The characteristics of this capturing method are: no selection, low capturing frequency, full automatic cataloguing, no quality control and a very low costprice per website. The Internet Archive started in 1996 as a private non profit enterprise, working to prevent the Internet as ‘a new medium with major historical significance from dissapearing into the past’. The webcrawlers of the Internet Archive do a new crawl of the www every two months. At the moment the Internet Archive comprises over 10 billion webpages, or 100 terabytes of data, which is five times the size of all the complete collection held by the Library of Congress.

The websites produced by public and commercial broadcasters are an implicit part of both the web projects of national libraries and of the large Internet Archive. They are captured for different reasons and in different ways: national libraries and archives regard this type of content as a part of the national cultural production. Depending on their collection policies, selections of broadcast websites will be stored and indexed. The crawlers of the Internet Archive copy broadcast sites periodically as an integral part of the global www. In both cases, the retention of this specialized content will be incomplete. Links to external sites might be broken and interactive navigation cannot always be retained. Robot crawlers are programmed to exclude copyrighted material and may skip pages that hold audiovisuals. The web sites of broadcasters that are collected as a result of these efforts will inevitably reveal omissions: missing pages, missing images, lost links, lost functionalities, no dynamic data access and a lack of continuity, due to random and irregular capture dates.

IV Web archiving in broadcast organisations

Within broadcasters, websites are generally produced by new media or internet departments. Their products are closely linked to the actual radio- and television production and may offer web formats of broadcast programmes and other footage, related information, and associated links. The websites are created and maintained from content management systems. New versions are generated every day. The content is automatically 'archived' in the sense that it is indexed and stored in the system. Although these systems are usually not designed to hold large digital archives, they turn into large digital archives all the same, including the proprietary- if any - archiving disciplines as to selection and indexing.

Generally, the more extensive and developed the organisations' webactivities are, the more internet departments have seriously contemplated long term preservation, sometimes leading up to special and well organized 'new media archives'. But among a lot of broadcasters, a formal, centralized policy for the deep-archiving of these materials is still a rarity. There's usually no clear organizational imbedding, and the area is simply too young and too alien for professional storage- and cataloguing protocols to have emerged. Although they may value its cultural significance - broadcasters may still have great difficulties to appraise deep-archived webcontent in regard to exploitation and re-usablity. On the other hand, selection dilemmas can hardly be underestimated, and may easily lead to a state of inertia, in which virtually every file and every version is retained.

In a sinister way the situation of scattered, copied and versioned materials reminds us of earlier days, when producers and journalistst used to sit on their own collections of films and tapes, because they were easy to have within reach and there was no-one in the company feeling responsible for long-term preservation anyway.

V The audiovisual archive as media-archive

The word 'media-archive' may refer to the merging of film- radio- and television collections, either organizational or digital. The term may also hint at the migration processes and at the online accessibility of the catalogues. A 'media-archive' may well be positioned in the heart of the digital production environment. But being a media-achive appearently does not automatically imply the task of preserving the organisations' new media-publications. A lot of archives do not seem too eager to get involved into the domain of web preservation just yet. On their way to becoming genuine 'media archives' however, these archives should seriously start to contemplate their role and responsibilities, and see to if and how the web might be included in their present actions. There is their mission is to preserve the organisations' legacy. But there is also the rapid emerge of generic, interactive multimedia materials, and the new delivery technologies such as web-TV, interactive television and multi-media home platforms. These developments will bring about a convergence of media that will cause traditional boundaries between genres, programme categories and distribution channels to vanish. The distinction between 'regular' audiovisual materials and multimedia objects - be it webcontent or production footage- will blur. Already, there are many editorial and content links between web sites and the broadcast radio- and television programmes. Born-digital or digitized broadcast materials are presented through the internet and often start a new life as part of a website. Newly produced audio en video materials are exclusively published on the Internet. Production departments and new media sections alike, will be producing these copied, versioned, intertwined, multimedia components, that together will make out the broadcasters' digital heritage. The old and the new collections of media-archives will inevitably be drawn into this whirling pool of digital objects.

Broadcast archives have built up infrastructures, standards and procedures to safeguard audiovisual products for future use and consultation. In selecting, cataloguing and preserving digital materials, these archives are already gaining practical experience, learning to understand the costs and the volumes, and developing technological expertise. Besides the archive's mission to actively preserve - taking into account the future convergence of media- there seem to exist the pure efficiency reasons, to consciously position the deep-achiving of webcontent closer to these archives.

VI Conceptual issues

Broadcasters' websites perfectly fit the description of the broader context audiovisual media: the audiovisual heritage, in parrticular where there is a reference to 'objects' and 'intangibles':
- recorded sound, radio, film, television comprising moving images and/or sounds
*- works**,** objects, materials and intangibles related to the audiovisual media, whether from a technical, industrial cultural, historical or other viewpoint.*
From a conceptual point of view, digital multimedia may even be regarded as audiovisual works. Apart from the element of 'lineair duration', Ray Edmundson's definition, in 1998 formulated as part of the "Philosophy of Audiovisual Archiving" still seems to stand:
Audiovisual works comprise reproductible images and/or sounds embodied in a carrier whose: a) recording, transmission, perception and comprehension requires a technological device and whose b) purpose is the communication of that content rather than use of the technology for other purposes.

But it is no use denying that the archiving of interactive multimedia appears a whole new and obscure discipline, that questions our professional standards and provokes our traditional skills. The nature of these media raises complex conceptual questions. The 'audiovisual archive paradigma' may state that "*the character of the audiovisual media and its products, are the first reference points for audiovisual archives*" but the point is that this very character becomes an enigma, once the materials are digizited. And, where it says that "*the audiovisual programme has to be viewed in its own right and not as an aspect of something else*" the emerge of dynamic, hyperlinked websites obviously wasn't foreseen.

The deep-archiving of webcontent raises two fundamental problems:
1. What is a digital object?
2. How to preserve a digital object in its original form?
These problems circle around the concepts of authenticity and integrity, concepts that presently dominate the worldwide discussion on the preservation of digital heritage. Authenticity and integrity can be seen as the basis of all theorectical and practical archival science. They imply that when we work with digital objects we want to know that they are what they purport to be, and that they are complete and have not been altered or corrupted.

In the past the building blocks of any broadcast archive were the discrete radio- and television works. The digitizalization is starting to severely influence the 'component parts' of our collections, that may pertain to:
1. Radio and television programmes
2. Audio and video materials
3. Digitized radio and television programmes
4. Digitized audio and video materials
5. Digital born radio and television programmes
6. Digital born audio and video materials
7. Webcontent.

The authenticity problem obviously increases on the scale from 1 to 7. If we cannot 'stabilize' a document or object, we cannot establish its authenticity. From edited, broadcast radio- and television programmes, catalogued and safely put away on our own shelves, we have documented source, treatments, contents and whereabouts. In the case of versioned production materials, that travel around along the production chain, the authenticity is already becoming harder to guarantee. In the networked, tapeless production environment, a lot of materials will have been reconstructed, compiled, summarized, abbreviated or otherwise manipulated, for different purposes. The 'respect des fonds', the principle of the original order, that implies that we are dealing with fixed entities, obviously cannot be maintained. In our secure, analogue broadcast archive environment, provenance and life cycle of audio and video materials were once relatively easy to establish. "Digital" authenticity problems are to us quite new, but will have to be conquered by all those, whose job it is to provide a reliable account of our past.

The integrity challenge, on the other hand, directly relates to the carrier-content principle that audiovisual archives have long been familiar with. This principle - the separation of the the carrier (physical film or tape) from the content of the work – has always been at the heart of our business. The very nature of av-media implies formats change and technological obsolescence, as well as the practicalities of access. Perception of audiovisual materials cannot but be influenced by the carrier and, in this respect, highly resembles digital content. This fact will not make digital preservation any easier. After all: converting a two-inch tape to a digital betacam is slightly different from migrating contents, look-and-feel and functionalities of an interactive website. But at least the organisational impact and the qualitative consequences of format progression are not completely unknown.

The dynamics of this progression however, will force us to adjust our traditional preservation strategies and find ways to maintain the technology rather than endlessly migrating the content from old formats to new. Many preservation standards will have to be reformulated. The 'loss principle', for instance, that a lot of archives have held on to (if there is any reason of form, content or external association to retain a particular item it must be preserved) becomes a useless precept in view of the digital flood that is heading our way. We – like all other digital archives and libraries- will have to redefine the concept of 'original' and still somehow manage to send certified 'copies' to go out into the world.

VII Concluding remarks

1. Media archives are already contributing a great deal to the dream of a well organized, infinite, digital multimedia library. A lot of their running preservation and digization projects are impressive. In transferring analogue films and tapes to digital formats, they focus on long-term storage requirements, on indexing and interoperability issues, on improving access.
But in developing into genuine media-archives they should however, consider becoming more pro-active in regard to the acquisition and deep archiving of webcontent. They should – in a more deliberate and explicit way– include these materials in their collection policies, for a number of reasons:

- Historical awareness and responsibility: webcontent is a part of the organisations' legacy and should be actively and consciously preserved for future generations.
- Efficiency: archives possess the professional attitude and already have laid out part of the necessary infrastructures and procedures.
- Convergence of media, genres and channels will eventually blur the bounderies between the traditional audiovisual collections and other digital multimedia content.

2. Convergence in the digital age also implies convergence of professional challenges, both conceptual and practical. Safeguarding the integrity and authenticity of our cultural heritage generates commonly felt problems and dilemmas among all types of archives and libraries. Audiovisual archives can now, and should then, team up with other institutes and take a much more active part in the international discourse that is taking place, on long term preservation of digital heritage. From a relatively isolated professional group - with specific technological and archival standards- the digital preservation issue offers audiovisual archives the chance to both benefit from, and contribute to the worldwide efforts that are currently being undertaken to come to terms with the digital domain.

October 2002, Annemieke de Jong