ICT Tools for Searching, Annotation and Analysis of Audiovisual Media

Alan Marsden*, Harriet Nock†, Adrian Mackenzie*, Adam Lindsay*, John Coleman†, and Greg Kochanski†

* Lancaster Institute for the Contemporary Arts, and Institute for Cultural Research, Lancaster University

† Phonetics Laboratory, University of Oxford

AHRC ICT Strategy Project report

October 2006

Executive Summary

1. This report concerns the use of ICT tools in research in the arts and humanities using speech, music, video and film in digital form, hereafter referred to as AV (audio-visual material).

2. The quantity of AV available to researchers is now massive and rapidly expanding, far exceeding the quantity of available print material in sheer number of bytes.

3. The main problem for researchers is no longer a paucity of AV but how to locate the material of interest in the vast quantity available, and how to organise material once collected.

4. Metadata and tagging continue to be important to facilitate search. Standards for metadata for AV do exist but are not yet widely adopted.

5. Content-based search is becoming possible for speech, but is still beyond the horizon for music, and even more distant for video and film. Mixed speech, music and noise is very hard to search.

6. Copyright protection hampers research with AV, and digital rights management systems (DRM) threaten to prevent research altogether.

7. Once AV has been located and accessed, much research proceeds by annotation, for which many tools exist. Systems for reuse and sharing of annotations are in their infancy, however.

8. Many researchers make some kind of transcription of AV, and would value tools to automate this process. For speech, such tools exist with important limits to their accuracy and applicability.

9. Full music transcription tools do not exist, but researchers can benefit from other sorts of visualisations, for which tools do exist.

10. Researchers could work more effectively with better knowledge of ICT. A common failing is not so much ignorance of how to use particular tools as a misunderstanding of the processes the computer carries out and the validity of its results.

11. In Section 1.3, recommendations are made concerning:
    i. provision of ICT infrastructure for arts and humanities research,
    ii. training for researchers,
    iii. copyright law and digital rights management (DRM),
    iv. resource development unlikely to receive commercial support,
    v. dissemination of expertise and examples in research on AV with ICT,
    vi. standards and commercial tools,
    vii. metadata and digitisation projects outside the research community,
    viii. management of researchers' private collections of AV,
    ix. deposit and sharing of AV, including annotations of AV.

Acknowledgments

We are very grateful to the following for their contributions to this survey: the Oxford 'Building a Virtual Research Environment for the Humanities Project' team: Ruth Kirkham, John Pybus and Alan Bowman; Bill Byrne, Stanley Chen, Colin Connolly, Peter Enser, Thomas Hain, Jing Huang, Giridharan Iyengar, Sanjeev Khudanpur, Roger Moore, Jiri Navratil, Mari Ostendorf, Christine Sandom, Andrew Senior, Sue Tranter, Phil Woodland, Ed Whittaker and many others for informal conversations. We also gratefully acknowledge the generous amount of time and information given by all of the participants with whom interviews are reported in Appendix C.

This project has been supported by a grant from the Arts and Humanities Research Council.

Contents

1 Project report. Audiovisual media, ICT tools, and humanities research

1.1 Introduction

Some of the most highly valued cultural forms in the west are stored in print form. Hence, much scholarly research focuses on what exists in printed form (e.g., the Bible, Shakespeare's plays, Descartes' *Meditations,* Beethoven's string quartets). This can give the impression that the humanities primarily refer to books and writing. Actually, much humanities research goes beyond print media. For reasons that go to the heart of their intellectual projects, scholars have been greatly concerned with ephemeral aspects of cultural materials such as speech, bodily movements, performances, and events. Visual, performance and mass media cultures generate transient materials, forms and processes that print represents poorly.

Media that record, store and transmit speech, music and moving images are roughly a century old. Electronic media and audiovisual recording technologies dramatically enlarge the horizon of cultural materials that can be analysed. Technologies that can handle sound, images and text in digital encoded form are just twenty-five years old. As well as static text and pictures, networked computers now deliver sound and video to the desktop. Even more recently, the audiovisual capabilities of a normal office PC open up possibilities for the easy use of non-print resources in many areas of the humanities. For example, historians may employ archive film or video footage of events, interviews, etc; artists, actors, musicians and others may study performances; linguists may wish to study the spoken language of such recordings, and so on. Libraries and universities around the world have been quick to explore the possibilities for making available such audiovisual materials to their researchers, and the internet allows users to access large quantities of audiovisual resources often without the need to go via established institutional providers.

However, humanities research has not always been able to quickly pick up on the enlarged possibilities of the universal media machine. The reasons for this are complex. Software to play sound and images has been mainly commercially produced. It has been designed to allow people to simply listen to or view media, without interrupting, repeating, searching or reordering it. By contrast, humanities research with such media typically relies on slowing down, comparing, collecting and sorting sounds and images in many different ways. Similarly, software for production of audiovisual material (sound editors, software sound and image mixers, video editors, video capture and encoders, etc) does not make it easy to analyse sounds and images. Typically, it makes it much easier to put sound and images together than to take them apart. Finally, in the last decade or so, there have been important research advances and large commercial investment in software systems that automatically transcribe, annotate or analyse sound and moving images. Even so, their application in humanities research is not straightforward. For instance, automated music genre analysis systems are technologically sophisticated and commercially significant. However, if identification of genre is an issue at all for current research in mu-

sicology, it is in the analysis of the process and concept of genre-association. If genre-identification software is to be useful in such research, it must be repurposed and perhaps unpicked to do more than simply assign a genre to a piece of music. Comparable illustrations could be made for video, film and speech research.

1.1.1 Scope of the report

This report explores the intersection between audiovisual media and digital technologies in the humanities as it stands in mid-2006. What can be done and what might be done definitely does not coincide with what is actually done. The report focuses on how research is carried out or could be carried out on materials that have already been recorded or captured in electronic form.

The scope of 'audiovisual media' for the purposes of this report is time-based audio and visual source material, as rendered through digital recordings or other capture processes. It excludes material based on still images, research material that exists primarily as a visual artefact (such as the image of a musical score), and materials which are primarily symbolic encodings or notations (such as encodings of a dance in Laban notation or a musical score). Although there are overlapping concerns, we exclude materials for teaching and materials used in a creative process (as in the performing, visual, and compositional arts).

It does not address other changes in research processes. For instance, the report does not explore how acquiring scholarly work through downloads or e-journals, or dissemination of results through electronic publishing changes the nature of research. Instead, we focus on problems and possibilities of working with primary materials such as recordings, footage or broadcasts. These problems and possibilities arise for contemporary scholars in many humanities disciplines. As much as possible, we have avoided addressing highly technical problems specific to single disciplines.

To highlight the most relevant points of intersection, we have adopted a simple generic model of humanities research using AV materials (Figure 1). The model views humanities research as a process of repeatedly accessing, searching, marking up ('annotating'), transcribing, analysing, and presenting

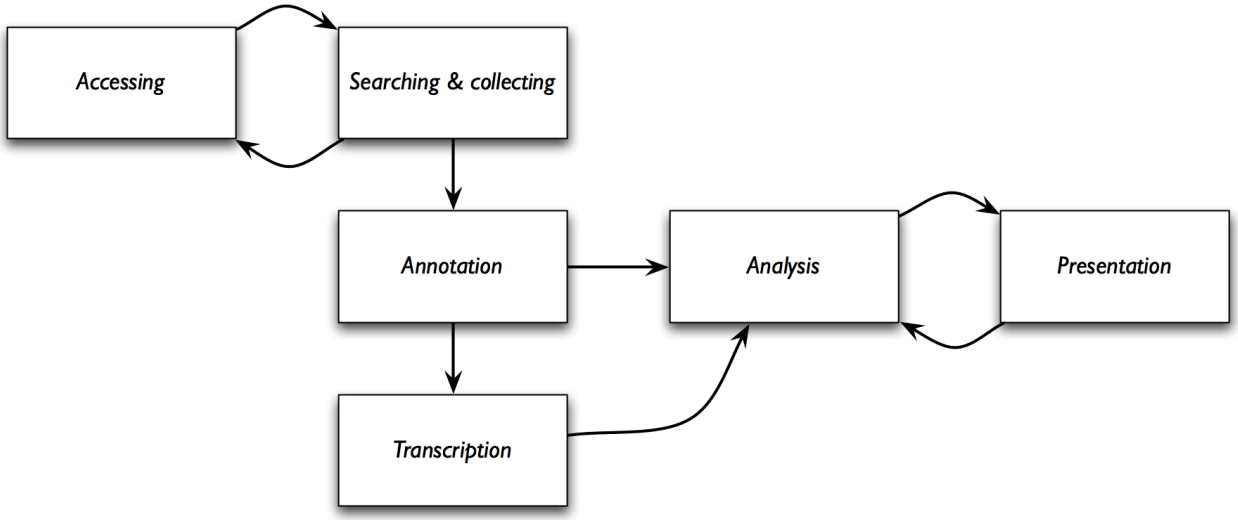

Figure 1. Schematic model of humanities research with audiovisual materials.

materials. As the figure suggests, the order of these operations varies. Scholars constantly cycle between different ways of working with audiovisual materials. Innovative research often combines them in unexpected variations or applies them to different materials.

1.1.2 Report website and project weblog

Resources in electronic form with hyperlinks for the examples and references contained in this report are available on the web as follows:

- An online version of this report: http://www.phon.ox.ac.uk/avtools (mirrored at http://ict4av.lancs.ac.uk/report)
- A weblog used in the process of gathering information for this project and including many more examples and links to recent developments: http://ict4av.lancs.ac.uk/

1.1.3 Other relevant reports

The recent British Academy Policy Review (2005) discusses factors which currently or potentially impact humanities and social sciences resources and access to them (see Section 4: Factors and Themes). Many of these factors — which include ICT advances, the Grid, access mechanisms, metadata, intellectual property and charging regimes — are directly relevant to the audiovisual resources considered in this report.

Relevant technology, copyright and privacy issues are discussed in the report of the EU-US Working Group on Spoken Word Audio Collections (SWAG, 2003).

A document from the British Universities Film and Video Council (BUFVC, 2004) highlighted the importance and varied uses of moving images and sounds and the difficulties faced by audiovisual archives. Key recommendations included collectively defining public sector audiovisual archive holdings as a distributed national collection and collectively asserting a public right of access to this collection for non-commercial use. More general archive-related initiatives such as the Archives Task Force and the UK Film Council's review of film heritage provision may also have consequences that are relevant.

The reader might find the following survey papers useful in addition: Goldman et al., (2005), Koumpis & Renals (2005), Lee & Chen (2005), Ostendorf et al., (2005). The survey of Goldman *et al* (2005) — a product of an EU/US (DELOS/NSF) working group on spoken word audio collections — also addresses policy issues relating to privacy and copyright and to the collection and preservation of spoken audio content.

1.2 Overview of the report

1.2.1 Organisation of the report

This main part of the report provides an overview and summary of conclusions. Appendices provide much more detail on different aspects of the model of research:

Appendix A: Accessing sources of audiovisual data. This section outlines the breadth and abundance of materials becoming available, and some of the difficulties that scholars encounter in making use of them. This includes the central problem of copyright law and Digital Rights Management.

Appendix B: Technologies for researching speech, music and moving image. This section summarises the major capabilities, possibilities, and lines of future development of digital technologies in humanities research with audiovisual media.

Appendix C: Current practices and expectations of humanities researchers. Based on a series of interviews and field visits to humanities researchers in a variety of disciplines, this part of the report outlines what researchers do and do not do, and what they would like to be able to do in their research.

1.2.2 Accessing audiovisual materials

The increasing volume of audiovisual materials is obvious. The quantity of data available exceeds the capacity of any library. Only a fraction of this is of interest to arts and humanities researchers, but it is no longer possible to identify a clear boundary between those materials which are of interest to scholars, and therefore should be preserved and made accessible, and those of no interest (if indeed it ever was possible to identify such a boundary). Researchers in the arts and humanities have a massive amount of material available to them, but it is less constant and less organised than the traditional text materials contained in libraries.

While there has been very significant growth in digital resources held by libraries and museums, often clustered around historical archives, accessing these materials, or the relevant parts of them, remains an issue. Often such collections are not easily searchable online, and the catalogues lack rich content descriptions. This can frustrate efforts to access material in a manner other than by the traditional identifiers of author, title and subject. In the space of ephemeral materials derived from popular culture, news media and general broadcast, an explosion of digital resources is occurring. Recorded music and film is mostly, sometimes solely, available in online formats. In consequence, the location of collections and repositories used by humanities scholars is shifting. Perhaps the most important access sites are no longer primarily institutionally managed. Instead, commercial services and user-produced archives and collections seem more important and relevant to much current scholarship. (Consider, for example, how significant Google has become in the everyday work of most researchers.) This situation yields greater certainty of access in some respects. For example, scholars are likely to have ready access

to a much larger collection of recorded music through an online music library such as Naxos than most research libraries hold.

However, the format of the audiovisual materials affects access in practical ways. Issues of fidelity are of diminishing importance with audio, but remain critical with video. What is sufficient quality for teaching and broad analysis is not sufficient for automated analysis or close analysis. Issues of fidelity depend crucially on what the researcher aims to achieve. Some formats, especially streaming formats, frustrate research processes, though they impose no absolute obstacle since stream-ripping software is widely available and increasingly used. Greater use of portable storage and the networking of media devices and computers makes it possible to access audiovisual materials in a wider range of contexts and in varying ways.

Copyright law and techniques of rights management are a much more significant factor in some areas of research. Copyright-restricted access is narrowing at a number of different levels. At a low level, Digital Rights Management (DRM) attempts to lock out any usage beyond the content provider's view of audiovisual media as entertainment. Access licences and pay-per-view schemes abound. Not surprisingly, a number of responses are emerging: new licensing schemes such as Creative Commons (2006) reserve some rights but with a general view that access should be open. However, access rights remains an important issue, and commercial usage of Creative Commons remains very small. Access can be a stumbling block for researchers who wish to work with audiovisual materials from established providers and, on the whole, is becoming a more serious problem.

1.2.3 Technologies — state of the art, gaps, obstacles

We have examined technologies of varying maturity, and do not limit ourselves to commercially deployed products or current ICT research. Some of these technologies were not necessarily developed for humanities research at all, but might be repurposed for scholarly work.

1.2.3.1 Searching and collecting

With vast collections of digital audiovisual material available, actually searching for and *finding* a resource can be a major barrier to research; if one is unable to locate a resource, one is effectively denied access. Nearly all practical, current, multimedia access depends on good-quality metadata for search. Content-based information retrieval is a field of active research, but for the most part has not yielded results such that the average, non-ICT-expert researcher can expect or obtain good results. Speech-based information retrieval is by far the most mature instance within content-based retrieval, and today's performance can be used as a rough indication of where video and music tools are headed within a decade. The performance of current automatic speech-recognition technology is at such a level as to make content-based retrieval practical for certain kinds of speech materials in restricted domains. Software tools or search engines to effect this do not currently exist outside the laboratory (though Blinkx TV (2006) is an on-line video search system that claims to perform speech search), but such tools can be expected to emerge in the near future.

For now, however, and certainly for music, video and film, searching is based upon catalogue data and other associated information; what one can find is a direct function of what metadata is associated with the resource. Until fairly recently, that was the sole responsibility of the archivist, but in recent years alternative strategies have come to the fore. Methods that search providers have utilised with audiovisual content include contextual (e.g., containing web page) and associated information (e.g., closed captioning information with a video). More recently, user-supplied metadata is starting to play a larger role.

1.2.3.2 Annotation

The metadata associated with a resource can be sufficient for locating a resource, but once a resource is found, there is often the need to associate finer-grained metadata with certain points within the audiovisual content. This potentially rich process is what we call annotation.

It is quite easy to annotate most forms of audiovisual material: video, speech, and audio annotation tools abound. Even so, it is often very time consuming to make annotations, so the challenge is to allow users to do so in a way that has some enduring value. One response is the development of standards to render annotations durable and facilitate their reuse by others. Important developments in this area are MPEG-7 and Annodex, but neither has as yet been widely adopted. Collaborative annotation systems are another means towards durability of annotations, by establishing a form of consensus, and they can also save effort by involving more users. On the other hand, user surveys indicate that ad hoc annotation happens all the time, sometimes involving pencil and paper, and the unpredictable nature of research means that this will always be the case.

1.2.3.3 Transcription

Transcription can be seen as an audio-only technology. As it is the process of fixing time-based events into a permanent medium, and video tends to be its own best document (what you see is what you get), speech and music have received the most attention and success for transcription purposes. As with speech search, which uses transcription as the basis for text-based search, we can look at the recent history of speech technology to get an idea of the future of music transcription.

Speech transcription, although continually improving in performance, has not fundamentally changed in fifteen years. There continue to be blocks to the dream of completely general speech transcription. One must choose a constraint, such as supporting a limited vocabulary, a single speaker, high training time or discrete speech (unnaturally separated, with pauses), in order to reach decent performance. On the other hand, while transcription of speech into accurate, properly formed and punctuated text might not be achievable, a transcription which provides information useful for some kinds of research is already possible.

Music transcription, now most commonly represented within the music information retrieval (MIR) community, faces similar blocks: performance is constrained by polyphonic streams, inaccurate tuning, and/or musical convention. Current technology is not remotely close to automatically transcribing any but the simplest monophonic music into proper music notation. On the other hand, as for

speech, transcriptions of other kinds which do show useful information are already possible, and the key challenge for MIR is to find those alternative views to note-based transcription which provide the most readily useful information.

1.2.3.4 Analysis

The location of 'analysis' in Figure 1 indicates our intended meaning for the term: while many of the tasks and processes of annotation and transcription are in some sense analytical, we mean here that part of research where the results of annotation and transcription are subject to the judgement and intervention of the scholar who seeks to extract useful information, draw lessons, and form conclusions. With respect to audiovisual materials, ICT tools play two distinct but possibly interrelated roles. The first might be described as 'microscopic analysis', where the tool makes explicit characteristics of or data about the material which is otherwise too small, too fast or otherwise hidden. The prime example is Fourier analysis and other systems which extract time-varying frequency information from an audio signal, important in the analysis of both speech and music. Another example important in music is the discovery of timing information to an accuracy of a hundredth of a second (or less). The second role for ICT tools is to facilitate navigation through audiovisual materials, especially multiple materials, multiple views of materials, or annotations or transcriptions in association with audiovisual materials. Tools make it easy for scholars to jump to specified locations in a source, to align similar materials, to see or hear them aligned, and to view or hear audiovisual material aligned with annotations or visualisations.

1.2.3.5 Presentation

Presentation refers to all the different ways in which digital technologies display or render different audiovisual materials apart from simply reproducing them. For instance, the timeline in a video editor or the waveform in a sound editor are presentations of images and sound respectively. Technologies that enhance presentation often summarise it in some way. Speech summarisation has been actively developed, and short textual or audio summaries from speech can be generated. Summarisation technologies music have also been a topic of research, for example presenting the salient features of a pop song in a few seconds, but have not yet been put to use outside the laboratory. Technologies that translate directly between spoken languages also offer new forms of presentation that could be useful in certain research domains.

Tools that generate visualisations of audiovisual materials are common. At the most simple, they display timelines of camera shots or audio events. They present information derived from audiovisual material in some graphic form, enabling overall patterns or structure to be seen, or assisting in the identification of points of particular interest. For instance, a timeline can be used to create a diagram of the formal structure of a piece of music. Existing sound and video editing tools can be repurposed for this. VJ (video deejaying) software allows many different video clips to be assembled, compared and ordered very quickly. This is a very active area of software development and use, and could well yield useful tools for research with collections of video material.

1.2.3.6 Integration

Technologies that integrate all the preceding research processes are few and far between. Even in the domain of speech processing and analysis, the area where analytical tools are strongest, there are few examples of 'integrated analysis environments' or packages of the kind that one finds in scientific software (for instance in bioinformatics, mathematics, statistics or engineering). The few tools that have begun to offer a fairly complete spectrum of analytical capabilities are large-scale, research driven initiatives. They are not currently very accessible to humanities researchers. The development of integrated analysis environments or 'knowledge studios' for humanities researchers remains on the distant horizon.

1.2.4 User experience and expectations

Humanities researchers whom we interviewed were treated as technology users for the purposes of this report. We sought to gather information on the 'life cycle' of audiovisual materials gathered for their research purposes, concentrating upon gathering resources, preparing data, analysis, and dissemination.

Audiovisual material generally falls into two categories: self-recorded or found material. In our observations, self-recorded material not only becomes a research resource, but can have a life as a work record, or take the form of research output or dissemination. Found (e.g., commercial or otherwise externally sourced) material usually only takes the form of a primary research resource, to be studied in and out of context. This situation is fairly natural: if the copyright lies with another party, it is often onerous for the researcher to obtain the rights for a small extract to appear in a research output. Both found and self-recorded material get classroom use.

Our user needs study interviewed 28 humanities researchers and several other technologists who work within the humanities. The research was carried out in three phases, starting with a general, cross-disciplinary study, and then moving progressively towards more specialised and audiovisual-specific research. General researchers were presented with screenshots and descriptions of certain exemplary projects garnered from the early phases of the technology review, and asked for their reactions. The specialist interviews focussed on specific needs and frustrations, and specific solutions were proposed or imagined in conjunction with the interviewers.

Self-recorded research sources were often based upon interviews, oral histories, or as documentary markers. There were few difficulties with recording equipment as it stands today. The common issue, however, was the vast amount of material collected and the limited time available to record and sift through it. As such, nearly all such interviewees wanted a solution for transcription of the material.

Found data runs into several potential roadblocks. The first is simply knowing where to look. As we have found, and as the report should demonstrate, not all large audiovisual archives are in obvious locations or maintained by the most obvious bodies to those accustomed to more 'traditional' textual scholarship. The second is that there can be access problems: although technical barriers to access are being lifted in the online world, not all of the most relevant archives are digitised or transparent to outsiders. Beyond simple access, access *rights* become terribly important in the digital world: DRM can cre-

ate difficulties from headaches and inconvenience to completely cutting off a legitimate line of inquiry on audiovisual material (e.g., automated signal processing and analysis on audio).

Once a data store is found and accessed, many find difficulty on the other side of the fence: there can be too much data for a single researcher to work on. A few researchers complained of coming across rich archives of video, but finding that manually tracking through for things that were personally interesting to them was too time consuming for the rewards. Again, transcription was an oft-requested *desideratum*, and implicitly demonstrates text's superiority for browsability over audiovisual material. Some researchers give themselves over to serendipity with found media, allowing broadcast media or online sources to open up new avenues for their research.

Once a particular piece of audiovisual content is chosen for deeper analysis, after an initial viewing/audition, a common first step is to develop some sort of timeline-based annotation. Although many ICT tools exist for this, many researchers are satisfied with making a table with notable time events, matched with other relevant notations, on paper. Those who deal with oral histories and other interviews cite making detailed transcriptions as a major effort and (often) expense. Further processing and analysis becomes much more individual to the researcher's personal methods and motives, but some researchers did show some interest in collaborative annotation (whilst expressing some doubt as to its technical or legal feasibility). Dissemination and other forms of sharing the results of research were similarly up to individual researchers. Those who had made use of ICT in doing so were generally comfortable with the tools available, since the tasks involved are familiar and well documented.

Finally, as technical experts speaking with humanities researchers, we noted some common misapprehensions about ICT tools and what they imagined the tools could achieve. Most of the problems arose from a complete trust in the infallibility of computers: that a computer could express uncertainty or offer a wrong answer flies in the face of most people's common understanding of computers. A few other problems came from those who were versed with the fact that computers *are* fallible: those researchers thought some operations were impossible with a given tool, when it was indeed possible, just obscured by the interface.

1.3 Conclusions and Recommendations

1. Network infrastructure and computing platform requirements for humanities research with audiovisual materials are growing and changing. Research with audiovisual materials typically requires higher network bandwidth, more storage, better audio and graphics processing capabilities, and display technologies than text-based research. Researchers are very interested and quick to pick up on devices and software that allow them to collect, view and search audiovisual materials. Poor or weak infrastructure thwarts experimentation with new research approaches. **We see a role for the AHRC in providing the (relatively modest) support for improved hardware and networks for humanities researchers.**

2. Many of the problems experienced by arts and humanities researchers working with AV are not purely technical, but involve broader issues. One of these is lack of knowledge and expertise, but the solution is not simply training in specific skills or with specific software tools (though these are important, and current efforts should continue, especially at the postgraduate level). Researchers in arts and humanities sometimes need broader knowledge of computing technology, its capabilities and limitations, and to be able to operate with statistical concepts of error and probability (as is common, for example, among researchers in some social science disciplines) in order to make proper use of ICT. **The AHRC should seek to foster knowledge of the capabilities and limitations of computing technology, and appropriate knowledge of error and probability, among arts and humanities researchers using ICT.**
3. Access and rights restrictions are important issues. Researchers are confused about their rights in dealing with audiovisual materials, and the law is indeed unclear in some respects. Clear guidance, where possible, to researchers about what they can and cannot do with audiovisual materials would be useful. We applaud the recent British Library manifesto on intellectual property (British Library, 2006a) and strongly recommend the AHRC to be engaged in public debate on this issue and to use its influence to establish rights of access to audiovisual materials for research. Digital rights management systems could, if widely adopted or even imposed in the way in which some companies propose, prevent effective research with audiovisual materials, even though such research would not harm the company's interests. **In view of the increasing importance of such materials for arts and humanities research, the AHRC must be involved in public debates and use its influence to prevent this. It is essential that the legal rights of access and the practical ability to access materials be maintained and promoted.**
4. The continuing digitisation of humanities-relevant AV resources and their exposure to the emerging AV search engines should make AV resources more appealing to researchers researchers for whom AV is not an essential primary resource. Such digitisation will happen anyway through the auspices of companies such as Google. **The AHRC should focus funds toward resource-creation projects that will *not* be covered by commercial advances.**
5. Many of the researchers interviewed could readily think of uses for AV data in their research, but researchers for whom AV is not an essential primary resource were often less than enthusiastic about pursuing these possibilities given the relative ease of locating and filtering text sources from their desk. **The AHRC should be active in promoting the research opportunities that access to AV resources allows**. Researchers currently using AV in their research could be explicitly recognised as 'early adopters' and facilitated to act as examples for others. Activities such as the AHRC Methods Network which engage with identified experts will assist, but it is important that the current focus on existing disciplinary communities does not allow important new areas to fall through the gaps. It will be important also to engage more directly with the developers of technologies, who might not readily envisage the applications in arts and humanities research to which their technologies might be put.

6. Commercial tools will develop rapidly and be usable for many purposes, often beyond those intended by the developers. However, with those tools, we expect there to be deficiencies and problems in applying those particular tools for arts and humanities researchers. Such problems include bibliographic inadequacies, lack of access to raw material, management of large quantities of diverse or unusual content, interoperation with other tools, black-box tools whose exact functioning is unknown, and access to intermediate results and material. Longer-term problems with closed formats and commercial software arise because there is generally no way to guarantee that such data will be readable or that such software will remain useable by future researchers. Although data format transparency has improved over the past few years with the widespread acceptance of XML as a carrier format, there are still dangers with closed formats or incomplete data output with an open format. **We recommend that the AHRC use and encourage the use of open, published data formats wherever possible.**
7. Similarly, we note that some AV digitisation projects and efforts to increase access are often undertaken with audiences other than arts and humanities researchers in mind. **The AHRC should monitor such activities as they are funded in order that arts and humanities researchers are consulted as stakeholders by at least some of these projects.** A particular humanities research need is the availability of as good and complete metadata as is possible during the whole archival, creation, and capture workflow: if that descriptive data is not captured, it is effectively lost forever.
8. Researchers face substantial problems in organising their own collections of materials so that they or others can use them. Content management has become a problem for individuals and groups of researchers. To date the AHRC has directed efforts at the generation of large and sometimes centralised collections of research materials while expecting researchers to cope as best they can with their own collections. However, private collections will continue to be an important part of many research projects in the arts and humanities. **The AHRC (perhaps through the Methods Network) should consider how individual researchers can be aided in generating and organising their own collections of audiovisual materials**. This should both facilitate their own research and facilitate the subsequent use of their collections by other researchers.
9. **The AHRC should encourage reuse of researcher-collected data via non-text-based support for browsing and exploration of AV deposited within archives.** For example, the Arts and Humanities Data Service (AHDS) could develop standards for the deposit of non-text materials, accompanied by appropriate metadata, and rich mechanisms for browsing and searching such deposits. It similarly could advise on or even require appropriate standards for the annotation of deposited audiovisual materials, and for the deposit of additional annotations of already deposited materials.

2 Appendix A. Accessing: sources and types of audiovisual media

There is a tremendous breadth of culturally interesting material in audiovisual form. A constantly growing proportion of it can be accessed via the internet. Under the rubric of *access,* we consider issues concerning the location of this material, its quantity, its nature, forms and format, and the problems of the availability of and right to use this material in digital form for research.

2.1 Digitisation

The transformation of pre-existing audiovisual material into digital form is largely outside the scope of this report. However, user interviews clearly indicate that this remains an important issue. The Arts and Humanities Data Service (AHDS) offers a good practice guide on *Creating Digital Audio Resources (*AHDS, 2006a). It 'aims to provide information and more specific technical guidance for those considering small or medium-scale audio digitisation projects. The guide is aimed at a non-technical audience and will be of interest to holders of analogue collections considering digitisation, managers who need enough information to plan resources for a digitisation project and those experimenting with or piloting digitisation on a small scale for research, teaching, promotion or creative projects.' (Plichta & Kornbluh, n.d.) gives somewhat more technical guidance.

Many archives have digitised some or all of their collections, or plan to do so, and this is often done in conjunction with a programme to make items available online. One such example is the Imperial War Museum's *Collections Online* (Imperial War Museum, 2006a). Another large UK digitisation effort is being led by JISC, 'the JISC digitisation programme' (JISC, 2006), funded with a £10 million grant from the Higher Education Funding Council for England. The program covers many resources, not just sound and moving pictures, but also archival sound recordings at the British Library (3900 hours) and Newsfilm Online (6500 hours) (JISC, 2005).

A published interview with the project manager of Newsfilm Online (eGovMonitor, 2005) reveals some of the complexities of these digitisation projects, which extend beyond the merely technical difficulties associated with choosing and converting data into formats that will be future proof and those associated with cataloguing. The interview reports that the project will have a licence to access the data in perpetuity, bringing access to hours of news film together with supporting metadata offering contextual information about a film as well as studio scripts and running orders and raw news feed for some time spans. Newsreel data will also be digitised. However, some of the data comes from third parties and, where copyright cannot be negotiated, part of the material will be fuzzed out and substituted with a caption that maintains ITN's commentary. More generally, decisions must be made about which data to include and which to leave out. The project comprises a steering group of academics as well as tech-

nical staff, and is conducting regular focus groups with higher education users to make sure that needs are met.

There is also interesting activity outside the UK: Google's efforts to digitise text collections are well-known, but their plans also extend to video collections as part of Google Video. They recently began posting the results of a joint digitisation pilot project aiming to make 'as much as […] possible' of the US National Archives public domain video content available online (News.com, 2006) (Google, 2006a).

There are many issues associated with the preservation of archived audiovisual material, both those which apply to all digital material needing to be preserved (e.g., Rosenzweig, 2003) and those which apply specifically to audio and moving image material (e.g., Besser, 2001). Interesting discussion about access and/or archive issues also arises in field specific papers, such as (Carson, 2005) and (Bignell, 2005), but these issues are beyond the scope of this project.

2.2 Quantity of data

Although digitisation projects continue to be important, it is now more common for research projects to have problems arising from too much rather than too little data. The UC Berkeley survey *How Much Information? 2003* (Lyman & Varian, 2003) gives some information on the volume of audiovisual information being created currently, including that outside archives. For example:

- World radio stations produce 320 million hours of radio broadcasting per year, of which 70 million hours are estimated to be original programming.
- World television stations produce about 123 million hours of total programming, of which 31 million hours are estimated to be original programming. The report also notes the growth in production of new movies and television, especially in developing countries.
- Their estimate of the amount of information recorded on the physical medium of film is somewhere between 76,000 and 420,000 Terabytes (1 TB = 1,000 Gigabytes). Note that this does not just include moving images, but also includes film-based content such as photographs and x-rays). (Lyman & Varian, 2003) also notes the beginnings of a move from film based cinema and TV into digital video (e.g. DVD) due to the lower costs of editing.

There are other significant sources of data, though these are flows rather than stored data: (Lyman & Varian, 2003) report that 'information flows through electronic channels (telephone, radio, TV, the Internet) are dominated by the information sent and received in telephone calls (including both voice and data on fixed lines and wireless), which if represented digitally would amount to 17.3 exabytes (17,300,000 TB) of new information.' Much of this information is ephemeral, but that does not prevent it from becoming a source for arts and humanities research. The capture and recording of eph-

emeral material is becoming increasingly common (for example, BBC radio's 'listen again' facility makes many radio programmes available online for a period).

(Lyman & Varian, 2003) also gives some indication of the amount of audiovisual data, reporting that approximately 370,000 motion pictures were made around the world from 1890-2002 and noting it would take 2108 years to play the entire universe of original film and video titles continuously. To put these quantities into perspective, the digitised version of the book collections of the US library of Congress would amount to 10 TB of information (19 million books and other printed collections). (Lyman & Varian, 2003) also comment on the often-discussed movement towards digital technologies and 'born-digital' data (i.e., data that is originally created in a digital format, rather than being converted to digital from some older recording format).

In the UK, some of this information is available to universities and colleges holding the appropriate Educational Recording Agency Licence or, for institutions holding BUFVC membership, through their Off Air Recording Backup Service. (Connolly (2004) summarises the rights situation from a modern languages and film studies perspective.)

2.3 Examples and sources of audiovisual data

Audiovisual data in collections around the UK and abroad include national or regional sound, film and television archives, television and radio company archives, newsreel archives, museum archives, stock libraries and academic collections, as well as numerous small collections held by local organisations, companies and private collectors. Some feel for the vast number of collections is given by the British Universities Film and Video Council's *Researchers Guide Online* (BUFVC, 2005), which aims to be 'the most detailed, specialised, accessible and up-to-date database … in the UK' focusing upon the subset of film, television, radio and related documentation collections in the UK: at the time of writing (early 2006) it currently lists 547 entries, including 118 core radio collections and 319 core moving image collections.

Examples of archives, selected relatively arbitrarily, include:

- The British Library Sound Archive (British Library, 2006b) currently contains over 550,000 hours of recorded sound (including categories classical music, drama and literature, jazz, oral history, popular music, wildlife sounds, world and traditional music, accents and dialects, sound effects). This is roughly 50 TB.
- The Presto media preservation project (Presto, 2006) estimated (based on 10 major broadcast archives) the total European holdings of broadcast material at 50 million hours in 2001, of which 10 million is film, 20 million is video and 20 million is audio (roughly 30,000 TB in total).

- The more specialised Imperial War Museum (Imperial War Museum, 2006a) holds 120 million feet of cine film, 10,000 hours of videotape and 36,000 hours of historical sound recordings (roughly 10 TB).
- Oxford University's Archive of Performances of Greek and Roman Drama (Wrigley, 2005), a specialist academic library, holds hundreds of video and audio recordings and thousands of other records of individual productions — photographs, programmes, reviews etc. — roughly 1 TB.

Such archives contain recorded speech of various forms, such as:

- Royal Academy of Arts annual dinner speeches from the 30s, 40s, 50s made by the BBC (British Library, 2006b) (BUFVC, 2002).
- The Black and Ethnic Minority Experience Project: recordings of interviews with over 100 Asian and Afro-Caribbean people who came to Wolverhampton (BUFVC, 2002).
- Recordings from Leeds University's 1950's Survey of English Dialects (British Library, 2006c).
- Millennium Memory Bank, one of the largest collections of oral history interviews ever collected focusing upon Britain (British Library, 2006d).
- Reminiscences of the composers by Debussy's stepdaughter and Prokofiev's son (British Library, 2006e).
- Performances by the Royal Shakespeare Company (British Library, 2006b).
- Margaret Atwood reading her poem 'The Immigrants', as part of The Poetry Archive, an online collection of recordings of poets reading their work (Atwood, 2006).
- Podcasts of recent sermons from St George's Church from Leeds (St George's, 2006).

Collections of recorded music exist also, and are increasing in size, but commercial interests and copyright mean that these are often not freely available.

- The recording company Naxos has made its entire output available on-line through a subscription service (165,000 tracks from 11,000 CDs) (Naxos, 2006).
- The UK's Joint Information Systems Committee's (JISC) collection 'Film & Sound Online' includes the Culverhouse Classical Music Collection of recordings (mostly 20th-century recordings of music from the 17th to 19th centuries (Edina, 2006).
- The Cylinder Preservation and Digitization Project of the University of California, Santa Barbara has put online the contents of 6,000 cylinder recordings dating from 1890 to 1930 (UCSB, 2006).

Short extracts from recordings are available online from many different sources; for example, it is now common for composers to make extracts available on their web sites. Other kinds of sound recordings are also available online, such as the sound of the Churchill tank starting up and moving off (Imperial War Museum, 2006b).

The situation for film and other moving images is similar to that for recorded music. Freely available non-commercial recordings do exist (e.g., All Go Margate (1970), one of a number of seaside re-

sort publicity films dating from the 1920s to the 1980s (South-East film and video archive, via moving history (AHRB Centre for British Film and Television Studies, 2005)). The Prelinger Archives (Prelinger Archives, 2006) holds over 48,000 'ephemeral' (advertising, educational, industrial, and amateur) films. CNN Image Source (CNN, 2006) contains CNN footage and makes it available to researchers for a price. Video upload sites such as YouTube (Youtube, 2006) are currently growing rapidly and comprise mainly 'home videos'. There are also collections of collections. The Moving Image Collections (MIC) (MIC, 2006) allows catalogue access to several dozen moving image collections.

Audiovisual data is also beginning to accumulate in new, digital institutional data centres and data repositories. Some of these are at the national level. For example, the JISC funded Film & Sound Online service provides access (including downloading) to film and video collections relevant to teachers and students and is hosted by EDINA, a JISC designated national data centre (Edina, 2006). There are also subject specific data repositories, such as those supported by the Arts and Humanities Data Service (AHDS, 2006b). This contains submissions such as the Designing Shakespeare collection, which includes a text database of production and review information, an image database of production photographs, a collection of video interviews with designers and a collection of VRML theatre space models (AHDS, 2005). Some institutions are developing their own data repositories (JISC, 2005). One such framework is the open source MIT DSpace framework (MIT, 2006a), which enables the submission, management and preservation of digital research material including (potentially) audio and video. The MIT iCampus OpenCourseWare initiative (MIT, 2006b) is currently archiving course materials in DSpace and another of the iCampus projects is investigating the audio recording of lectures for later search and retrieval (MIT, 2006c).

Audiovisual data is also increasingly being made available through web and pay-per-view services. With increasing broadband uptake, film and TV-over-PC is becoming more popular. Relevant data extends as far as the performing arts e.g. the UK Theatre Network plans to produce the world's first pay-per-view theatre (announced October 21, 2005), building on technology trialled for film downloading and sports to allow users to log on and view the latest play either live or recorded (OpenPress, 2005).

Other sources of audiovisual data include individual user generated content, which is increasingly born digital. Lightweight and easy-to-use technologies such as webcams, computer microphones and digital video cameras such as those in mobile phones as well as new editing technologies make data collection and manipulation considerably easier for individuals than in the past. Once collected, some of this data is made available on the Web in the form of pod casts (audio blogs), vlogs (videologs), moblogs (comprising content posted to the Internet from mobile devices, in this case devices supporting audiovisual capture), posted as art or uploaded to sites such as Google Video (Google, 2006b) (to be discussed later).

The Google Video National archives effort represents one of several on-line archives of material which has fallen or been given to the public domain. One such example is The Open Video Digital Library (OpenVideo, 2005), a publicly accessible digital video repository. The repository was developed in

part as a testbed for video retrieval researchers but also to serve the practical needs of the public for an open collection of video: the collection spans categories including documentary, educational, ephemeral, historical and lecture (Marchionini & Geisler 2002).

2.4 Technology and formats

Audiovisual data is held in such archives and collections in a variety of formats. For example, the British Library supplied the following statistics for their recorded sound collection (Robinson, 2005):

- 163 hours on cylinders.
- 26,000 hours on coarse-groove (78rpm) discs.
- 175,000 hours on vinyl (33rpm) discs.
- 53,000 hours on open reel 1/4 inch magnetic tape.
- 41,000 hours on audio cassette.
- 213,000 hours on CD.
- 49,000 hours on other digital carriers.

A typical small video collection is the Oxford University *Archive of Performances of Greek and Roman Drama* (Oxford 2005a) which includes an audiovisual collection consisting of 250 videotapes and perhaps a hundred CDs and audio tapes.

Some material is available only in streamed formats (e.g., the Naxos Music Library), which carries issues of reliability and fidelity. Other material (particularly film and sound) can be available in compressed formats which might or might not lose information important for a research project. The Cylinder Preservation and Digitization Project takes the useful approach of making its material available in both compressed restored formats and in a high-resolution raw (unrestored) format; each format is likely to be more suitable for different kinds of research projects. The use of analog base magnetic tape (audio and video tape) has also decreased as digital storage has increased. The production and sale of retail audio CDs has declined, whilst DVDs have achieved the fastest market penetration of any recent technology innovation.

The issue of digitisation formats is much less critical than it was five years ago. Many recent desktop machines have the storage, memory, processing power, and network bandwidth to deal with 'consumer-grade' audio and video without a problem. Still, there are some age-old principles that remain, and appreciating them can help any person who works with audiovisual resources:

- When compressing video or audio, there will always be a multi-way trade-off between quality, size, and computational complexity. It is possible to squeeze high-quality video into a relatively small file size (or bandwidth requirement), but the trade-off is that the computer processor has to work much harder to view the video (and even harder to compress the video).
- When capturing audio or video, especially for archival purposes, capture and store in as high quality as possible. Once compressed in order to save disk space, it is impossible to

get the lost information (and audio/video quality) back. Note that compression algorithms are tuned to do minimal damage to the audio/video, as perceived by humans. Automated analysis and annotation tools may use different features, so their performance may be impaired by compression.

- Choose a codec carefully for compatibility and longevity. Although some vendor-specific schemes for compressing and decompressing video may show short-term gains and boast of high compatibility through market share, it is always best to go with open standards, backed by multiple vendors and implementations, such as the MPEG family of codecs.

Although it is not a general principle, relying on streamed media for scholarship serves nearly no one today. Although it makes some sense in a mobile data scenario, it is retrograde for researchers, who frequently need to hop around the material, slice it up, and focus on small sections.

2.5 Platform survey

As previously mentioned, desktop computers are currently capable of viewing high quality audio and video. They can store and manage a moderate collection of audiovisual resources. Desktop digital video editing is well within modern computers' capabilities. Future computing capacity could go to even higher quality video, and/or viewing multiple streams. Typical current query-by-audio- or video-content algorithms, however, run roughly equivalent to real time, so would have a very difficult time being applied to large collections. For this, collaboration, a pooling of resources, and/or looking towards Grid technologies might be the way forward. Bringing a massive sharing of computing resources to bear on shared, batch processing of a known corpus of audiovisual resources could be the next step forward in creating online audiovisual resources.

The other trend, parallel to the increased computing power on the desktop, is the increased computing power in mobile devices. This points to a more significant change. Audiovisual resources have begun to accompany researchers throughout their work and personal lives (e.g. using a video iPod to store film collections).

2.6 Availability

Not all audiovisual archive material is readily available for research purposes. Firstly, not all of the data held in archives is catalogued: (Sandom & Enser, 2003) report that many film archives have large and growing backlogs of items for which there are no content description. The Presto Project, which examined archives of broadcast material, found 'the content is unavailable to the general public and often unavailable even to national archives and educational institutions. Much of the content is unique, e.g. master material that cannot be allowed to circulate generally, and all of the content has rights issues'. (Presto, 2006) (But see comments on the Creative Archive later.)

Some archives maintain private catalogues, others make catalogues available via the World Wide Web but require appointments to be made and travel to the archive to view resources. Others offer some or all of their collections online, discussed below under digitisation. Where data is catalogued, it may not be catalogued consistently across archives (Sandom & Enser, 2003), although there are certainly efforts in this direction, such as the Open Archives Initiative (OpenArchives, 2006) which formed the basis of The Open Language Archives Community (Simons & Bird, 2003).

2.7 Access rights

Access to material can be restricted for commercial and copyright reasons. This is particularly true of film and recorded music. The Naxos Music Library referred to above, for example, is available at a cost. The JISC collection Film & Sound Online is currently only freely available until 31 July 2007. The Cylinder Preservation and Digitization Project makes its materials freely available. Although there are currently no copyright restrictions on the original material, copyright does apply to the restored digitizations. However, controls are waived for non-commercial use.

Looming large over all issues with audiovisual content in the humanities is the development of digital rights management (DRM): multimedia content owners want to protect their content and profits, and for distribution of digital-only content, insist upon some form of anti-copying technology. When it comes to research, especially as aided by data analysis tools, this becomes a grave concern and a major obstacle.

DRM is the general term for a variety of technical solutions (reinforced by legislation) designed to allow the rights-owner of content to determine how a consumer may use the content. In the case of digital audio, rights may be limited to listening to the content on a limited number of computers and/or associated compatible portable devices. In some cases, the rights may be time-limited, such as when the rights to listen are tied to a monthly subscription. The status of DRM'd content (that is, content protected by some digital rights management system) that consumers pay for now no longer resembles the ownership that people have been accustomed to in the case of physical media. DRM'd content, lacking any tangible form, is now licensed or leased, ultimately subject to the will of the rights owner.

The general technical method for implementing DRM is to encrypt the file and tie the encryption key to the content-purchaser, the computer, and/or the date. A specialised, trusted application on the computer or portable device has the ability to decrypt the file and play it. No other applications may do so. This causes difficulties. DRM'd content lacks compatibility with data analysis methods. No applications other than those 'trusted' by the DRM scheme provider have access to the decrypted content: such applications typically do not offer the analyses that researchers need, and even if they did, the algorithms are unknown and undocumented, so the results are of uncertain value to researchers. A DRM scheme provider will want to know what an application does with the decrypted content before granting it 'trusted' status, so research applications cannot be trusted *a priori* because the researcher cannot

know in advance all that is to be done with the decrypted content. Data analysis programmes are shut out of working with DRM'd content directly. Cumbersome workarounds may be possible but are often impractical for large amounts of content.

The current trend toward strong protection for intellectual property will likely harm many research activities in the humanities and social sciences. It already restricts research in many areas. The especially damaging trend involves the combination strong intellectual property laws with Digital Restrictions Management (DRM) software. The problem comes about because DRM software typically is not written to allow the 'fair dealing' exceptions that are allowed by copyright law. Thus, in practice, researchers are losing their rights to access data.

For instance in the UK, Section 29 (1) of Part I of the Copyright, Designs and Patents Act 1988 as amended (2003) states 'Fair dealing with a literary, dramatic, musical or artistic work for the purposes of research for a non-commercial purpose does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement.' This clause should allow access to films, music, and documents for a wide variety of University research. However, the authors of this study are not aware of any DRM software that actually implements Section 29 (1). Broadly speaking, DRM software is written to make it hard for consumers to pirate the content, and researchers are – incidentally – treated as pirates.

DRM software is not just a technological trend, it is enforced by law. It is illegal to circumvent any copy protection scheme, and even illegal to construct or possess devices and computer programs that will be used to circumvent DRM. An example case where the High Court upheld the law was (*1) Kabushiki Kaisha Sony Computer Entertainment Inc (2) Sony Computer Entertainment Europe Ltd (3) Sony Computer Entertainment UK Ltd v (1) Gaynor David Ball & 6 Ors,* [2004] EWHC 1738 (Ch), 19 July 2004. Such a law is required by treaty obligations. Consequently, one cannot circumvent DRM software to gain access to protected content, not even for allowed research purposes. Even if it were legal to break DRM protections in pursuit of a fair dealing use, one could not legally possess the required tools.

DRM technology is also converging with the efforts of a group called TCPA (Trusted Platform Alliance), which aims to build hardware to allow strong control of what software can access what data. Other names for this are TC (Trusted Computing) and NGSCB (Next Generation Secure Computing Base). While this technology has its benefits, if adopted it will allow content providers to specify how software will display their product. For instance, a supplier of music could (and presumably would) require that Windows Media Player shall send the music only to the speakers and no where else. It would then be difficult and illegal to analyze the music with other software to understand the details of the musical performance.

Another related technology that comes under the general heading of DRM is CPRM (Content Protection for Recordable Media). CPRM is a mix of hardware (within the disk drive) and software (in an application program), and it aims to encrypt data as it is written to the disk in order to control the copying of sensitive data. This technology has been implemented since 2004 (CyberLink, 2004). If

broadly implemented, it would check any disk accesses against rules provided by the content providers. The press release makes it clear that the technology is intended for controlling the playing of videos on DVD. Such a technology would be a severe problem for someone in film studies or someone who was studying advertisements. Likely, such a researcher would need to collect excerpts (or adverts), but the technology would prevent him or her from copying parts of the DVD.

We want to emphasize that there are many technologies under the general heading of DRM. By the time this report is read, the details may change. However, there is a strong economic incentive for entertainment companies to implement DRM so we expect that DRM will not disappear. Conversely, there is no significant economic incentive for companies to preserve the 'fair dealing' exceptions specified in copyright law, so researchers cannot expect unhindered access. We note that the same problem also arises in using copyrighted materials for instructional purposes. The only practical solution seems to be an exception to DRM legislation that would allow the use and possession of circumvention tools for non-commercial research purposes.

2.8 Altered rights management

The most well-publicised response to the assertion of strict copyright is the Creative Commons model (CreativeCommons, 2006) and related UK variants. These legal instruments allow authors to reserve some, rather than all, rights (e.g. the right to benefit if the material is reused commercially). They provide a compromise between the extremes of copyrighted and public domain. This model has spurred the development of sites storing audiovisual data for certain kinds of reuse (particularly for creative, non-commercial purposes), such as the BBC Creative Archive and The Freesound Project. There exist other similar sites, for example the Internet archive movies section (InternetArchive, 2006). As with The Open Video Digital Library and similar projects, these archives may also prove useful for technological development, as well as stimulating creative artistic works.

The BBC Creative Archive (BBC, 2003a) was first announced in 2003 and is intended to increase licence payer access to the archives through the Creative Commons inspired creative archive licence (BBC, 2003b); clips can be downloaded for non-commercial use, stored on PCs and edited and shared. Releases so far include clips from Radio 1 and 1Xtra and BBC news; future releases for the 2005-2006 pilot programme include the subjects of science and nature. Other participants in the group include the British Film Institute, the Open University and Teachers' TV (BBC, 2003a). The Freesound Project (FreeSound, 2006) is an Internet-based project supporting the free exchange of sound effects through a website which allows anyone to participate by contributing or downloading files. Sounds are made available under the Creative Commons 'sampling+' licence, which allows most uses of the sounds provided the source is acknowledged.

Public-good archives of data themselves encounter obstacles. BBC's experiment in free downloads of Beethoven's nine symphonies was heavily criticised by classical music labels. They maintained that the BBC was diluting the commercial value of their products (Seltzer, 2005).

3 Appendix B. Technologies for researching speech, music and moving image

This survey considers current technologies for some loci of in Figure 1 (p.2): accessing, searching and collecting, annotation, transcription, and analysis. Consideration will be given not only to technologies currently in use, but also to those which are the subject of research or development and likely to come into use by 2010. We adopt the following classification to indicate the current stages of development of the tools discussed:

Category 1: Mature project

Category 2: Usable but still under development

Category 3: Technical demo

Category 4: Proof of concept

Category 5: Lab experiment

The reader should note that classifications assigned are not exact. For example, many state-of-the-art research technologies could be described as falling into categories 3, 4 and/or 5.

3.1 Other sources of information

This report gives an indicative rather than comprehensive survey of current tools and technologies. Other lists, surveys or collections of tools include the following.

- A useful survey of Internet free/shareware tools for voice analysis is maintained at http://www-users.york.ac.uk/~dmh8/dmh_pevoc4.htm, although it is not clear how actively it is being maintained (last update March 2005)
- A comprehensive list of speech analysis (and transcription) software is maintained at http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html, some of which also handles video.
- Another useful survey examining freeware, shareware and commercial digital speech processing tools is Gonet and Święciński (2002).
- Among its advertised resources, the International Computer Music Association maintains a 'software library' (in fact a collection of links to sources of software) (see http://www.computermusic.org), but some of this is now out of date.
- PALATINE, a Subject Centre of the Higher Education Academy also maintains a set of links to music software as part of its 'Directory' (http://www.lancs.ac.uk/palatine/directory.html).

3.2 Searching and collecting

Research in principle starts with some kind of searching and collecting of materials. The search for relevant materials often relies on previous analysis, annotation and sometimes transcription. There is no absolute point of origin for searching since almost every search relies on a prior categorisation. Searching and collecting take very different forms, and the technologies needed vary widely.

3.2.1 Searching the spoken word

The most widely used and highly developed search systems work with text, and so searching spoken word collections often relies on previous annotation, transcription or content analysis (topics covered in later sections) to derive text from, or associate text with, the spoken word.

3.2.1.1 Transcript search

Systems supporting the free-text querying of textual transcripts are now ubiquitous and similar systems exist for searching speech by querying the time-aligned transcripts automatically derived by speech-to-text systems. Surprisingly, with relatively little engineering ingenuity, the errors and lack of punctuation in these automatically derived transcripts have little impact upon the effectiveness of the search performance once the error rate falls below an often quite achievable rate of around 40%, at least for the types of data and tasks investigated to date (for more details and possible explanations, see Allan, 2003). However, difficulties may arise in scenarios where the transcription system vocabulary is inadequate to capture all the words in the content, because speech clips containing a word which does not exist in the automatically derived transcripts can never be found: such a problem is more likely to arise with dynamically evolving collections such as daily news (Hauptmann, 2005), rather than static archives, though this is not an absolute rule. There has been research into techniques for handling this problem, including techniques for searching secondary phonetic transcripts: a query term which falls outside the system word vocabulary is (hopefully) located by searching for its pronunciation in the phonetic transcripts (Logan et al., 2003; Amir et al., 2002). (In fact, some companies adopt the use of phonetic searching as the primary search mechanism and claim that this gives better results than a word level search, but this has been hotly disputed by researchers due to the lack of publicly available results.)

3.2.1.2 Browsing via metadata

Metadata (generated manually or automatically) can be added to indices of various types, analogous to but more flexible than those found for books. Thus, a user might choose to browse only segments corresponding to a particular speaker or those that have been associated with particular named entities such as people, places or locations. (There remains considerable art in designing interfaces for supporting efficient browsing through such metadata. Section 3.6.1, ‘Summarisation’, is of relevance here.)

Tools of these types have appeared in digital library scenarios since the 1990s (e.g., similar ideas, although less powerful technology, appear in the Princeton digital library (Wolf & Liang, 1997) and the Kansas digital library (Gauch et al., 1997)). Automated speech indexing technology is also beginning to appear in Web search tools, discussed in Section 3.2.4, 'Tools for locating AV on the web'.

Companies offering category 1 tools include Aurix (2006, was 20/20 Speech), Scansoft Dragon MediaIndexer and Scansoft Audio Mining (Scansoft, 2006), BBN Audio Indexing System (BBN, 2004-6a), Nexidia/Fasttalk (Fasttalk, 2006) and also Autonomy (2006) which offers both speech search and video search solutions (via Softsound and what is/was Virage). Ted Leath's First Year PhD Report (Leath, 2005) makes a brief comparison of a subset of commercial products and research systems including BBN rough'n'ready, FastTalk/Nexidia and ScanSoft MediaIndexer. (Companies exploiting such technology for Web AV search are discussed in Section 3.2.4.) There are also numerous research projects in categories 3-5 that are attempting to develop more sophisticated systems, and some of which are discussed in Section 3.7, 'Integration'.

Another issue affecting search systems for multilingual spoken word collections relates to the handling of users who generate queries in languages different to the collection material. One solution is to include a human with appropriate language skills in the search process; the technical community is also attempting to address this problem under the Cross Lingual Information Retrieval umbrella. Typical solutions include translation of the query to match the language(s) in the collection, translation of the collection to match the query language, or representation of both query and language using some intermediate or 'interlingua' representation. Much of this work is in categories 3-5 and has addressed the broadcast news domain, although there has been limited category 4-5 work addressing more conversational and emotional speech as part of the MALACH project (Oard et al., 2002).

3.2.2 Searching for music and sound

It is important to be clear whether in 'searching for music' we mean finding information about where music can be located or actually gaining access to the music itself, equivalent to the distinction between a search yielding the bibliographic information for an article or the full text of the article itself. The possibilities for the former are currently much greater than the latter.

Most searches, whether within a collection such as the Naxos Music Library (Naxos, 2006), or within a database such as Gracenote (2006), depend on metadata such as title, composer or performer. The Naxos Music Library gives access to the actual streamed sound (for subscribers), while the Gracenote database gives catalogue details of CD recordings. In both cases the metadata is restricted and problematic, and based very closely on information provided with CDs. For example, titles of pieces may appear in a different language from the original composition.

In a few cases, the metadata associated with a recording is expanded to arbitrary tags, for example in the Freesound project (2006). The efficacy of this depends entirely on the usefulness of the original tags, generally collected through some collaborative process (see Section 3.3.3, 'Collaborative annotation'). An approach which does not depend on explicit tagging is to assume that the text in prox-

imity to a reference to a music or sound file is usefully associated with that music. Thus a search for 'Beethoven' might yield music files which have 'beethoven' in their title, or in the text of links referring to them, or which is linked from pages with 'beethoven' in the title. Google's American site (not the UK one) offers a 'music search' facility which is called up when a search is recognised to refer to an artist (see Google, 2005). Searching for 'Beatles' for example gives access to specific searches related to each of their songs, but searching for 'Beethoven' does not currently trigger any music search. Altavista (2006) allows search results to be restricted to audio files and will indeed give access to recordings of Beethoven's music in response to a search for 'Beethoven'.

There has been considerable interest in searching for music using sound rather than text as the search term, called 'query by humming'. While there have been a number of experimental systems (category 3-5), some of them available on the web (e.g., NYU, n.d.), none has reached the stage of a usable tool. There are very significant technical issues to be addressed before this can be achieved and questions about the degree to which it would ever be a simple-to-use and effective tool (Pardo & Birmingham, 2003). If humming is not a good interface for finding music, an alternative is demonstrated in Muugle (Bosma et al., 2006) (category 5), which provides an on-screen music keyboard on which a user may play a query, which is then matched against the database. Input from a MIDI keyboard is also possible.

3.2.3 Searching video and film

Databases that list information about films and television shows are now common on the web. These databases rely on simple genre classifications, the names of producers/directors, publication information, subject keywords and sometimes other content-related information. For instance, the Internet Movie Database (IMDb, 2006) (category 1, finished product) provides reviews, plot summaries, much technical production information and sometimes trailers for over 800,000 films and television series (July 2006). Because volunteers have added so much information about plot summaries and characters to the database, it can be used to find films and television programs by subject, genre, etc. For computer gaming, projects such the Open Directory Project (2006) (category 1) or Games-db (2006) (category 1) offer something similar but without much of the production related information. Apart from player reviews, they focus on 'cheats' – instructions on how to play games more easily.

For current television content, new content alert systems based on program schedules provide automatic notification of broadcasts that fit certain criteria (e.g. MeeVee (2006) (category 1) or Radio Times (2006) (category 1)). The BBC has announced its commitment to making 1 million hours of television and radio searchable and available online. The BBC Programme Catalogue (BBC, 2006). (category 3) allows 75 years of broadcasting to be searched.

On the horizon of searchability, systems that bridge different media are under active development. For instance, search engines that range across television and web contents have been designed (e.g., Miyamori et al., 2006).

However, these systems really do not actually search the content of film, video or broadcast. As in the case of the spoken word resources, they still rely on previous cataloguing, annotation or transcription. Even the most advanced video upload sites such as Yahoo! Video (2006) (category 1) require submitters to supply the keywords used for indexing and cataloguing the clip. Web search engines will probably index video at a fine grained level as collaborative annotation techniques develop (see Section 3.3.3, 'Collaborative annotation'). This topic is discussed further at the end of the following section.

3.2.4 Searching for AV on the web

There are already some established methods of accessing audio and video on the Web, some targeted specifically at researchers and the arts/humanities. These include portal efforts such as the BUFVC's Moving Image Gateway (BUFVC, 2006), which collects links to websites involving moving images and sound and their use in higher/further education, and HUMBUL, which includes categories such as Modern Languages – General, Sound/Audio (HUMBUL, 2006). Another example is the work of OLAC (The Open Language Archives Community), which has extended the Open Archives Initiative infrastructure in order to support creation of virtual digital libraries comprising distributed language resources: community efforts not only support standardised resource discovery (including spoken audio and also associated tools) but also recommend best practice for resource creation (Simons & Bird, 2003; Goldman et al., 2005).

More generally, online suppliers of content for online and offline viewing are rapidly increasing in number. For example, iTunes allows the download of video content (e.g. TV shows or media company pod casts) for transfer to a video-capable iPod and TiVo now supports transfer of content recorded by the TiVo to an iPod or PlayStation portable. Video coming through these mechanisms have established charging mechanisms, some per month, some per content unit.

We distinguish these offerings from emerging Web search tools aimed at locating AV. Most of these fall into categories 1 or 2, and they share many similarities with search engines for text. One of the earliest of these was Speechbot, a general Web deployed tool for audio indexing speech recognition transcriptions. Speechbot supported many of the functions now familiar for text-based searching, allowing free text, advanced or power searches and produced a results list displaying a number of items comprising a 10 second long errorful transcription around the located (and highlighted) query terms, the ability to play the corresponding 10 second extract and the date of the recording. Speechbot is now unavailable due to the closure of the Compaq Cambridge Research Lab (US), but in the past couple of years a number of similar services have emerged. Many of these emerging services have been released as test or beta versions for audio and/or video, and changes to the functionality offered by any one site are appearing almost weekly at the time of writing. For this reason, we describe typical functionalities rather than describing specific systems in depth.

Some tools crawl the Web for audio and video made openly available on websites. For example, podscope offers the ability to search audio blogs and pod casts, as does Blinkx (2006). Truveo offers a similar service (Rev2.org, 2005).

Some tools support the search of video or audio submitted by users. For example, podscope allows users to submit content (Price, 2006a), while Google Video operates the Google Video Upload program (Google, 2006c), whereby video and optionally a transcript are submitted to the system.

Some tools index content legitimately provided by media companies and archives. For example, blinkx has major deals with ITN and Fox News Channel (net imperative, 2006). Yahoo! Video and Truveo accept videos through media RSS (Rev2.org, 2005), Google video operate a Premium Program for major producers (Google, 2006d) and also have a pilot project with the US National archives (News.-com, 2006; Google, 2006a).

The tools perform the search in different ways. Some rely on metadata associated with videos, such as web page captions or user uploaded transcripts (the current version of Google video may fall into this category). Others extract closed captioning or use speech-to-text technology to allow more precise indexing as discussed earlier, returning results which play from the point of the first-matching query term (e.g. TV Eyes (Price, 2006b)). Most of the sites described offer services in English; services are also appearing in Mandarin (e.g., Blinkx, 2006) and Arabic (TV Eyes, 2003).

The business models of these companies are still evolving. Services such as blinkx appeared to be inserting advertisements into searchable content (net imperative, 2006). Others offer premium fee-based services e.g. TVEyes (Price, 2006b).

3.2.5 Content management systems

Video and audio are 'large media'. On the web, in film and video databases, on DVDs, in legacy collections of video and film, there is no shortage of film footage or television content. Many scholars collect large amounts of this material on their own computers and on portable storage media. While professionally curated online archives usually have extensive catalogues and indexes, personal collections of audiovisual materials sometimes suffer from lack of organization.

At one level, the folder and directory structures available on desktop computers allow virtually any material to be organised. However, others means of organizing audiovisual materials are available. Most music and video player software such as iTunes, xmms or windows media player embodies some idea of bookmarks, libraries or playlists. Annotation software often includes file management features, sometimes for thousands of files. Dedicated personal information management software such as DEVONthink (Devon Technologies, 2006) handles multimedia and text files equally. Commercial media management software such as CONTENTdm (Dimema, 2006) and Retrieva (2006) offer more sophisticated ways of organizing contents. Some software attempts to automatically index still images and text files added to it. Their search capabilities only use tags and metadata for sound and image files.

3.3 Annotation

In the context of time-based media, annotation associates extra information, often textual but not necessarily so, with a particular point in an audiovisual document or media file. In humanities research, annotation has long been important, but in the context of sound and image, it takes on greater importance. Rich annotation of content is required to access and analyse audiovisual materials, especially given the growing quantities of this material. Annotation software for images, video, music and speech is widely available, but it does not always meet the needs of scholars, who annotate for different reasons. Sometimes annotation simply allows quick access or index of different sections or scenes. Annotation has particular importance for film and video where Annotation is sometimes used for thematic or formal analysis of visual forms or narratives. At more fine-grained levels, some film scholars analyse a small number of film frames in detail, following camera movements, lighting, figures, and framing of scenes. Annotation tools designed for analysis of cinema are not widely available. Most video analysis software concentrates on a higher level of analysis.

3.3.1 Annotation and standards

There are many different approaches with regards to standards in annotation. There are several well-known metadata standards applicable to humanities research, such as library standards like MARC and Z39.50, and other, broader standards like the Dublin Core. These are useful standards, but are dominated by the resource-level approach; most similar metadata standards describe content on the level of an entire entity within a library. This level of metadata is very useful, but does not satisfy the requirements of annotation as described above: the standards do not have robust models for marking points *within* the content.

MPEG-7 is an ISO standard (category 1), conceived in 1996, and finalised (in its first versions) in 2001-2002. It is intended to be a comprehensive multimedia content description framework, enabling detailed metadata description aimed at multiple levels within the content. It is worthwhile to go into a little detail on the standard and what it might offer to humanities researchers.

A key to understanding MPEG-7 is appreciating the goals that shaped its conception and the environment in which it was born. It was conceived in a time when the World Wide Web was just showing its potential to be a massively interconnected (multi-) media resource. Text search on the web was beginning, and throwing into relief the opacity of multimedia files: there was then no reliable way of giving human or computer access ‘inside’ a multimedia resource without a human viewing it in its entirety. Spurred on by developments in the general area of query by example (including query by image content and query by humming), it was thought that MPEG could bring its considerable signal processing prowess to bear on those problems.

Along the way to the standard, people discovered that the problem of multimedia content description was not all that trivial, nor could it rely wholly upon signal processing. It had to bring in higher-level concerns, such as with knowledge representation and digital library and archivist expertise. In

doing so, the nascent standard became much more complex, but had the potential to be much more complete.

The standard, as delivered, has a blend of high- and low-level approaches. The visual part of the standard kept closest to MPEG's old guard, concentrating on features unambiguously based upon signal processing and very compact representations. The newly created Multimedia Description Schemes subgroup (MDS) brought in a very rich, often complex set of description structures that could be adopted for many different applications. MPEG-7 Audio took a middle path, offering both generic, signal processing-inspired feature descriptors and high-level description schemes geared towards specific applications.

Technically, MPEG-7 offers a description representation framework expressible in XML. Data validation is offered by the computationally rich, but somewhat complex XML Schema standard. Users and application providers may customise the precise schema via a variety of methods. There are numerous descriptive elements available throughout the standard, which can be mixed and matched as appropriate. Most significantly, it allows for both simple and complex time- and space-based annotations, and it enables both automated and manual annotations.

Industrial take-up and generally available implementations of MPEG-7 have been inconsistent at best so far. The representation format offered by MPEG-7, however, seems to be one that would serve arts and humanities research very well. It is agnostic to media type and format. It is very general, and can be adapted to serve a variety of different applications. Despite its flexibility, it is far more than a 'guideline' standard: it has very specific rules for ensuring compatibility and interoperability. If someone were to invent a framework serving the arts and humanities research community for its metadata needs, it would resemble MPEG-7, at least conceptually.

A fine-grained approach to the problem of re-using annotations relies on developing shared standards for annotation. Standards for annotation of video content have been developed. e.g. Annodex (2006), (category 2) is an open standard for annotating and indexing networked media, and draws to some extent upon experience gained from MPEG-7. Annodex tries to do for video what URL/URI (i.e. web links) have done for text and images on the web. That is, to provide pointers or links into time-based resources of video on the web. The Metavid project (2006) demonstrates Annodex in action on videos of U.S. Congress.

3.3.2 Manual annotation

There are numerous tools (and formats) for creating linguistic annotations, many catalogued by the Linguistic Data Consortium (2001). (According to the LDC, "'Linguistic annotation" covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions – audio, video and/or physiological recordings – or it may be textual. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, "named entity" identification, co-reference annotation, and so on. The focus is on tools which have been widely used for constructing annotated linguistic databases, and on

the formats commonly adopted by such tools and databases.') Some of the analysis tools mentioned earlier also support annotation, see e.g. Gonet and Święciński (2002) or the long catalogue of tools listed by Llisterri (2006). There is also the open source Transcriber tool (2006) and numerous other commercial solutions for more general transcription of digital speech recordings, such as NCHSwiftSound (2006). These tools fall variously into categories 1-4.

For video, a typical video annotation tool is Transana (category 1) developed by WCER, University of Wisconsin (2006), which allows researchers to 'identify analytically interesting clips, assign keywords to clips, arrange and rearrange clips, create complex collections of interrelated clips, explore relationships between applied keywords, and share your analysis with colleagues.'

Annotation of music associates non-textual information with the original data more often than is the case for other media. For example, scholars needing to know where the beats come in a piece of music might associate a sequence of MIDI data with an audio or MIDI stream. Essentially the task remains the same: to associate some symbolic information with points or segments of the audiovisual medium. In the case of multi-channel or multi-track data, it is possible that annotations might be applied to separate channels or tracks, but we have found no instances of this. The kinds of annotations which researchers wish to make range from structural or quasi-semantic labels (e.g., 'first subject' or 'mellow') to technical/analytical data (e.g., harmonic analyses, key or tempo) to the identification of small-scale events (e.g., specific notes or drum beats). These annotations can be attached to time points in the original stream, or to segments. In the latter case the annotations might or might not form a hierarchical structure (with segments contained within segments) and might or might not containing overlapping segments. Annotation tools are unfortunately rarely explicit about which of these kinds of annotation are supported. Lesaffre, Leman, De Baets & Martens (2004) discusses some of the theoretical issues around musical annotation.

Though not intended explicitly for annotation, music-editing or music-composition software can have annotation capabilities or be repurposed to perform annotation tasks. One example of the use of commercial sequencer software is Tanghe, Lesaffre, Degroeve, Leman, De Baets & Martens (2005), who used Cakewalk Sonar (by Twelve Tone Systems) to annotate the drum beats in extracts of sound recordings. Two MIDI tracks were added using the software, one indicating where the beats came and the other indicating each percussion stroke and the kind of instrument (bass drum, snare drum, etc.). An advantage of using the software was that the MIDI track could be played either with or without the original audio track, allowing the user to check by ear whether or not the percussion strokes had been correctly identified and correctly timed.

Tools intended for annotation of speech could be used also for annotating music, and while Lesaffre et al. dismiss these as not suitable because they do not support the kinds of annotation required for music, WaveSurfer (category 1; Sjölander & Beskow, 2000, 2006) has been used both directly and as a basis for specialised music-annotation tools. Similarly, generic tools for annotating video or audio, such as Project Pad (Northwestern University, 2006) (category 2), can be used for annotating music. Perhaps the most highly developed specific music annotation tool is the CLAM Music Annotator

(MTG, 2006; Amatriain, Massaguer, Garcia & Mosquera, 2005) (category 2). This software allows different kinds of annotations to be attached to time points or to segments, and different kinds of annotations can attach to different segmentations. Annotation types can be defined using an XML schema, and software elements can be added to automate some annotation processes (see Section 3.3.4.2).

3.3.3 Collaborative annotation

Annotation of audiovisual materials can take a lot of time, and even if material has been annotated by one researcher, the problem remains of how any other researcher can make use of the annotation. It is therefore not a surprise that projects have investigated sharing the effort and the results. We see much activity in this area, and some promising early ideas. Annotation for the purpose of finding audiovisual material seems successful, but we have not seen anything like the sophisticated and consistent analysis that would be needed to write even a basic film or book review.

Simple collaborative annotation of audiovisual materials is now common on the web. Sites such as Google Video (Google, 2006b) (category 1) or Youtube (2006) (category 1) partly rely on tags supplied by contributors. Producers and consumers of audiovisual material such as photographs, speech, sound and music, or video tag them with keywords. These keywords then become searchable via web search engines or through subscription mechanisms (e.g., a user who subscribes to content tagged with 'Star Wars' will receive notification whenever anything tagged with 'Star Wars' has been added to the database). While people often choose very generic keywords, and the keywords often apply to large video files, the tags and keywords are clearly useful. There is a synergy between the descriptions supplied by different users. For example, one may annotate the style of the image, and another marks the presence of a street sign. Combinations of the annotations supplied by users allow database-driven websites such as flickr.com and youtube.com to provide reasonably powerful and selective search capabilities, more informative than one would expect from any single set of annotations. Currently, commercial video uploading and downloading services are growing rapidly, and they offer increasingly sophisticated annotation features (e.g. Viddler, 2006, category 3). However, by and large, the annotations only describe the most obvious features, which limits the searches that can be done.

A number of projects have attempted to design and construct collaborative software environments for video annotation. In collaborative video annotation, a number of people can work on the same video footage. Efficient Video Annotation (EVA) (Volkmer, 2006) (category 2) is novel Web tool designed to support distributed collaborative indexing of semantic concepts in large image and video collections. Some video annotation tools such as Transana (WCER, 2006) already exist in multi-user versions. Another approach to collaborative annotation is to set annotation up as a game in which you get annotations generated as a side-result. See, for example, The ESP Game: Labeling the Web (Carnegie Mellon University, 2005a), which is a collaborative gaming approach to image annotation that is described in more detail in (von Ahn & Dabbish, 2004). The same team has a later game, Peekaboom (Carnegie Mellon University, 2005b), which helps in generating labels for segmented images, useful for

computer vision, for example. The designers claim to have generated over 10 million descriptive words for one million images.

A different approach is taken by the application mediaBase (Institute for Multimedia Literacy, 2005) (category 2), which requires some manual annotation or tagging of any media file put in the system. However, after this initial tagging, it encourages 'rich media authorship' as a way of investigating relations between different media components. MediaBase publishes resulting compositions on the web, and they can be altered, edited, revised or added to by others. The goal of MixedMediaGrid (NCeSS, 2005) (category 4), an ESRC e-Science funded project, is to generate tools and techniques for social scientists to collaboratively analyse audio-visual qualitative data and related materials over the Grid. Certainly, these tools and techniques could be used in the humanities too. MediaMatrix (category 2) developed at Michigan State University (2005) is a similar online application that allows users to isolate, segment, and annotate digital media.

Similarly in music, a number of projects have suggested processes of collaborative annotation to allow researchers to pool effort and benefit from each other's annotations. Project Pad (Northwestern University, 2006) is designed explicitly to allow teams (envisaged as students, but they could be researchers) to share annotations. Collaborative annotations of music in education have been reported, but none in research. A collaborative music-annotation project intended for research has been set up at Pompeu Fabra University, using either the CLAM Music Annotator or a Wavesurfer-based client to a web portal (Herrera et al., 2005), but there is as yet little evidence that it is accumulating a large set of annotations.

The BBC also has a project in this area, with the aim that listeners will progressively annotate recordings of radio programmes (Ferne, 2005). This is an internal BBC research project, but a public launch has been mooted. Interestingly, this project uses a Wiki-like approach, allowing the public to edit existing annotations, including viewing histories and reverting to previous versions, but with the underlying assumption that there is a single canonical annotation.

Already established and an everyday part of music on the web, but not really a research tool, is the Gracenote database of CD tracks (Gracenote, 2006). The database supplies annotations for media players to supply information about artist and title which is not recorded in the electronic data on an audio CD. Publishers of CDs can supply the original information to Gracenote, but many CDs were published long before there were media players on computers, let alone before the Gracenote database existed. Commonly, when a media player finds there is no information on the database for a CD, the user is invited to supply this information, which is then sent to the database. Thus Gracenote is effectively a global collaborative annotation tool. However, in the area of classical music recordings, it is notoriously inaccurate, largely because the categories for the database ('Artist', 'Song', etc.) do not map clearly to the commonly significant details of a classical composition (e.g., is the 'Artist' the composer, the soloist, or the conductor?). Research uses for the database are therefore likely to be confined to popular music.

3.3.4 Automatic annotation

An alternative response to the time-consuming nature of manual annotation is to automate part of the process. Clearly, different kinds of annotations present different levels of difficulty in automation, and it is in the simple and explicit partitioning of audio, in particular, that automatic annotation has had the greatest success. The challenges of more 'semantic' levels are much greater, though some projects in this area have had a degree of success, particular with respect to music.

3.3.4.1 Audio partitioning

The goal of audio partitioning systems is to divide up the input audio into homogeneous segments and (typically) to determine their type. The class types considered may vary by application but a typical partitioning might distinguish pure music, pure speech, noise, combined speech and music and combined speech and noise (Tranter & Reynolds, 2006). The resultant partitioning may provide useful metadata for the purpose of flexible access, but such partitioning is also an important prerequisite for speech-to-text transcription systems (e.g. it enables the removal of audio that might otherwise generate transcription errors) (Gauvain & Lamel, 2003). For some applications, a more knowledge-based partitioning and filtering may be applied, such as removing advertisements and other audio segments that are either not of interest to the end user and/or are likely to degrade automated system performance downstream. Such technology falls into categories 3-5 and is typically available from companies and/or research labs with interests in speech-to-text transcription.

3.3.4.2 Music

The past decade has seen the birth and rapid growth of the field of Music Information Retrieval (MIR), fed in part by the interest of music businesses in technologies to facilitate user interaction with large databases of downloadable music. While 'query by humming' (see Section 3.2.2, 'Searching for music and sound') was an initial impetus to this field, more research has recently been directed at what are effectively various kinds of annotations of music. Some of these are concerned with partitioning (e.g., note onset detection or segmentation into broad sections) and some concerned with richer information such as tempo, beat, harmony and tonality, and various kinds of similarity or classification. Two well developed tools for MIR are Marsyas, by George Tzanetakis (Tzanetakis, n.d.; Tzanetakis & Cook, in press), and M2K, by J. Stephen Downie and others (Information Systems Research Laboratory, 2005), which functions within the D2K 'Data to Knowledge' framework of the US National Centre for Supercomputing Application.

The achievements of recent MIR research are best shown in the results of the MIREX competition (MIREX, n.d.) associated with the international conferences on Music Information Retrieval (ISMIR, n.d.). The 2005 competition had ten categories: 'Audio Artist Identification', 'Audio Drum Detection', 'Audio Genre Classification', 'Audio Key Finding', 'Audio Melody Extraction', 'Audio Onset Detection', 'Audio Tempo Extraction', 'Symbolic Genre Classification', 'Symbolic Melodic Similarity', and 'Symbolic Key Finding'. The 'audio' competitions used recorded sound as the raw data, while the 'symbolic' competitions used MIDI files. The best audio systems typically performed with accuracies of 70–80%, and

though the key finding approached 90% accuracy, this is still well below the level at which such software would produce reliable results with real saving of effort if details of individual cases are important. The best symbolic systems interestingly performed at similar levels of accuracy, despite the much lower complexity of the input data. Other tasks on symbolic data, on the other hand, such as 'pitch spelling' (i.e., determining a note name and accidental for each note such as 'C sharp' or 'D flat') can be performed with levels of accuracy of greater than 98% (Meredith, 2006), promising useful research tools. Most MIR software falls into categories 3-5. Only Marsyas has become sufficiently widely used to take on the status of category 2 (released, but not yet finished, software), but its use is currently as a toolkit for MIR research rather than a tool for musicologists.

It would be a mistake, however, to think that MIR research will not assist musicological and music-analytical research. While it is true that tools which automate the typical tasks of music analysis are, as yet, not in prospect, MIR tools do produce a wealth of potentially useful and interesting data about musical sound of a somewhat different nature (e.g., measures of acoustic roughness, and various kinds of correlations). With a change of focus by music analysts (and a certain amount of re-education, since the acoustics and mathematics involved are not part of the general knowledge of music analysts), these tools promise novel and fruitful areas of research which focus on the analysis of music as sound rather than music as notated structure.

3.3.4.3 Video

A video can be partitioned into shots. A shot is an uninterrupted segment of video frame sequence of time, space and graphical configurations. For the last decade, many research projects have been working on automated video partition of footage into shots, topics, and face recognition (particularly in news video processing). Some of this research has led to commercial products. Some of these systems use manual annotation to start with, and then automatically annotates and indexes any related video materials. For instance, the Marvel video annotation system (IBM, 2006) (category 3) demonstrates the ability to generate semantic and formal labels from television news footage. Marvel builds statistical models from visual features using training examples and applies the models to automatically annotate large repositories. Other projects seek to generate topic structures for TV content using TV viewer's comments on live web chat rooms (Miyamori et al., 2006).

3.4 Transcription

Transcription is typically applicable only to audio within time-based multimedia. More technically, as it is a process of writing down events in a canonical form, it applies to events that are transitory and constrained. As such, music, dance, and speech are the most commonly transcribed sources of those within the project's remit. Automatic general video transcription makes little sense in the near-term because it essentially requires a model of the whole world. With constrained worlds, some transcription is possible, and there has been some automatic video 'understanding' of sports on video as well.

3.4.1 Speech-to-text transcription

Speech-to-text (or automatic speech recognition) systems aim to convert a speech signal into a sequence of words. Progress in the field has been driven by standardised metrics, corpora and benchmark testing through NIST since the mid-1980s, with systems developed for evermore challenging tasks or 'speech domains': developing from the domain of single person dictation systems to today's research into systems for the meetings and lectures domain. A brief history of speech (and speaker) recognition research can be found in Furui (2005a).

Some of the differences between speech domains can create additional difficulty for automatic systems. For example, speech from the lecture domain has much in common with speech from a more conversational domain including false starts, extraneous filler words (like 'okay') and filled pauses ('uh'). It also exhibits poor planning at higher structural levels as well as at the sentence level, often digressing from the primary theme. An evaluation in 2004 reported state-of-the-art transcription systems to achieve a 'word error rate' (a measure of system accuracy which incorporates word deletions, insertions and substitutions) of 12% for broadcast news in English, but 19% for Arabic. For conversational telephone speech, the figures were 15% for English and 44% for Arabic (Le, 2004). The effect of differences in the manner of capture of the audio is illustrated in the figures from an evaluation in 2005 for meetings and lectures (in English), where the error rates were 26% and 28% respectively when speakers had individual headset microphones but 38% and 54% in the case of multiple distant microphones in the meeting room (Fiscus, 2005).

Development of a system for a new speech domain or application ideally builds upon a large amount of manually transcribed in-domain 'training' data in order to build a speech transcription system tailored to that domain (often of the order of hundreds if not thousands of hours for state-of-the-art systems (Kim et al., 2005). The level of accuracy of the transcriptions need not be perfect: techniques have recently been developed to handle less than perfect transcriptions such as closed captions (Kim et al., 2005): technologists report that up to a 5-10% word error rate can be handled in a single transcript or multiple transcriptions of different reliability exploited (Phil Woodland, personal communication). Where sufficient adequately transcribed data cannot be made available for financial or other reasons, as much adequately transcribed in-domain acoustic data as is feasible is obtained – which will sometimes be none – and models from a 'similar' domain are adjusted or adapted in terms of their acoustic, vocabulary or word predictor components in order to match the new domain as well as possible. Vocabulary and language model (word predictor) adjustments can also be made based upon in-domain textual information such as transcripts, textbooks or other metadata where available.

There is a computation time versus accuracy trade-off: a real-time system will typically perform less well than a 10-times-real-time (10xRT) or even unconstrained system, but the degradation will vary with situation. Similarly, memory constraints can affect things. State-of-the-art systems typically use hardware beyond that of today's average desktop. (The word-error rates for English speech referred to above were achieved with a constraint of 10xRT and 20xRT respectively (Le, 2004).)

It is important to note that speech recognition systems developed for one domain cannot, in many if not most situations, be employed as a black box that can handle any domain: even speech from the same domain that differs from the 'training' data may be problematic (e.g. speech from previously unseen broadcast news shows in Le, 2004). There exist components of the system which are 'brittle' or sensitive to such changes: the system has been trained to recognise certain types of speech and, whilst it may perform quite well on those types of speech, it may perform badly on speech which is 'different'. Such differences may include (but are not limited to):

- channel differences, such as speech which is recorded over the telephone versus speech which is recorded using a headset microphone;
- individual speaker differences, including accent, vocal range;
- style of data, whether conversational, dictated, produced and carefully pronounced (as in broadcast news);
- vocabulary.

There exist system adaptation techniques to compensate for such differences to some extent (e.g. Gales 1996), but despite significant progress in this area the development of systems which are robust to differences in data is a key research goal at present (Le, 2004; Ostendorf et al., 2005).

Systems have also been developed for some domains in many other major European languages e.g. the LIMSI-CRNS spoken language processing group has developed broadcast news transcription systems for French, German, Portuguese and Spanish in addition to English, Mandarin and Arabic (Gauvain & Lamel, 2003). Mention should also be made of the recently-started DARPA Global Autonomous Language Exploitation (GALE) program (see Linguistic Data Consortium, 1996-2005), which is developing technologies to absorb, analyse and interpret huge volumes of speech and text in multiple languages: as part of this, projects such as AGILE (Autonomous Global Integrated Language Exploitation, involving multiple sites including the University of Cambridge and the University of Edinburgh) are developing combined speech-to-text translation systems that can ingest foreign-language news programmes and TV shows and generate synchronised English subtitles (Machine Intelligence Laboratory, 2005). (Such technology is becoming commercially available for certain scenarios, though at a cost hinted to fall in the band of 'hundreds of thousands of dollars' (US dollars): IBM has recently developed the TALES server system that perpetually monitors Arabic TV stations, dynamically transcribing and translating words into English subtitles; the video processed through TALES is delayed by about four minutes, yielding an accuracy of 60 to 70% compared to a estimated 95% human translator performance (PC Magazine, 2006).

Differences in speech transcription performance across different domains mean that speech transcription tools fall into development categories 1-5 depending upon the difficulty of the domain. For example, desktop systems for dictated speech-to-text and desktop control are readily available (e.g. Dragon NaturallySpeaking), as are systems for constrained domains such as medical transcription (e.g. Philips SpeechMagic supports 23 languages and specialised vocabularies). The Microsoft SDK can be freely downloaded and used for the development of speech-driven applications and is supplied with re-

cognisers for US English, simplified Chinese and Japanese (Microsoft, 2006). All of these tools fall into categories 1-2 but will perform well only in certain situations.

State-of-the-art speech-to-text systems are typically made available through joint projects with universities or commercial organisations such as Philips and Scansoft. These tend to fall into categories 2-5. For the enthusiast with time to spare, the HTK project offers downloadable software that will let you build a reasonable word-level or phonetic-transcription system and now offers an API (called ATK) for building experiment applications (HTK, n.d.); SPHINX-4 (Sphinx, 1996-2004) is an alternative, and there are many other tools of interest, such as the CSLU Toolkit (Centre for Spoken Language Understanding, n.d.). These tools fall into categories 4-5.

Church (2003) presents a chart showing that speech-to-text transcription researchers have achieved 15 years of continuous error rate reduction and we might wonder what the future holds. At present, the accuracy of current systems lags about an order of magnitude behind the accuracy of human transcribers on the same task (Moore, 2003; David Nahamoo quoted in Howard-Spink, n.d.). Moore has estimated that it would take a minimum of 600,000 hours of acoustic training data to approach a 0% error rate using current techniques, which he also estimates to be a minimum of four times a typical human's lifetime exposure to speech!

3.4.1.1 Speech-to-phonetic transcription

Researchers have also investigated automatically extracting textual transcriptions comprising a sequence of sub-word units (e.g. syllables or single sounds referred to as 'phones'). This task has not been as heavily researched in recent years, but has relevance to search and indexing applications since such subword transcriptions often form the basis for techniques for searching for 'out-of-vocabulary' query words with which the word level transcription system is not familiar. New words appear in the news every day (e.g. '9/11' suddenly entered our vocabulary) and may not appear in the basic speech to text system vocabulary, so could not be recovered in a straight word transcription search. (A discussion of techniques for handling OOV ('out of vocabulary') queries in spoken audio can be found in Logan et al (2003).) Recent work examining phonetic transcription includes Saraçlar et al. (2000) and Saraçlar & Khudanpur (2004). Phonetic transcription tools typically fall into development category 5 and exist within universities and research labs, though only for specific phone sets and not necessarily in forms which are easily packaged. The NICO toolkit (KTH, n.d.) also supports development of a neural network-based estimator of phoneme probabilities, though this would probably be of interest only to hardcore enthusiasts.

3.4.1.2 Transcription with video

As in some other problem domains, there is some convergence in research based on audio and video. Audiovisual speech-to-text systems, which combine information about the movement of the lips and possibly a wider region of interest around the mouth with audio information, have been found to improve over audio-only speech-to-text in certain conditions (e.g. noisy conditions and/or large vocabu-

lary dictation tasks) (Potamianos, 2003). Category 2 tools of this type are under development for constrained domains such as finance and within-car use.

Allowing the combined use of audio and video was also found to improve the segmentation of stories on video, relative to purely speech transcript-based approaches for most systems at TRECVID 2004 (Kraaij et al., 2004); multimodal information retrieval systems can also outperform speech based retrieval systems, although speech-based retrieval contributes most of the performance to date (Hauptmann, 2005). Multimodality can also be usefully exploited in presentation e.g. the CueVideo system offers the end-user a choice between presentation formats such as visual-only storyboards (slideshows of key frames without audio) and moving storyboards with audio, allowing them to select the most appropriate presentation mode for the video content in use (Amir et al., 2002). All of these research areas are still at quite preliminary stages, with the exception of audiovisual speech recognition work, and fall mostly into categories 3-5. However, it seems likely that solutions which make use of multiple rather than single information sources, where this is an option, will prove most successful in the future.

3.4.1.3 Time-alignment of speech and text

A convenient property of the most popular (statistical) approach to speech recognition is that the same algorithm used for speech to text transcription can be used to time align a word level transcript (e.g. a script) with the corresponding speech signal, associating each word associated with its start and end time in the audio signal. (In a robust system, the algorithm may be lightly modified to allow for errors in the script, e.g. Chan & Woodland (2004) and for distracting nonspeech audio such as music or other background noise.)

3.4.2 Transcription-related annotation of speech

The speech-to-text transcriptions as discussed above have historically comprised an unpunctuated and unformatted stream of text. There has been considerable recent research into generating 'richer' transcriptions annotated with a variety of information that can be extracted from the audio signal and/or an imperfect word level transcript. Such annotations may improve applications which involve presentation of transcripts (e.g. user reading of results returned by search systems), but may also improve downstream processing (e.g. machine translation). Areas of interest include:

3.4.2.1 Punctuation and structural information

There has been investigation into automatically generating punctuation as well as into generating speech-specific structural information such as marking interruption points, edit regions and boundaries of sentence-like units. Much of the latter work fell under the umbrella of the DARPA EARS program, under the structural metadata task (Liu et al., 2005).

3.4.2.2 Speaker-related information

Associated tasks include speaker detection and tracking (identifying a number of speakers and grouping utterances from the same speaker, although absolute speaker identities remain unknown), speaker

identification (determining who is speaking at a particular time), speaker verification (determining whether a particular speaker corresponds to a claimed identity) and tasks related to speaker localisation (e.g. in meeting scenarios). Examples of such work include the summary paper by van Leeuwen et al (2006) and (Tranter & Reynolds, 2006).

3.4.2.3 Named entity extraction

The task involves annotating transcripts to mark word sequences corresponding to items such as proper names, people, locations and organisations, or dates and times. The BBN Identifinder (BBN Technologies, 2004-6b), which is a category 1 named entity extractor that has been quite widely used in the technical community.

3.4.2.4 Topic-related information

Tasks investigated include the detection of topic boundaries in a stream of data, clustering of related segments of data, the automatic detection of later occurrences of data relating to the story of interest and story link detection tests to determine whether two given stories are related. As this description suggests, these tasks make most sense for news data which comprises a sequence of stories although there has been related work for conversational speech such as that in the MALACH project. (Allan, 2001 includes a summary of topic-detection and tracking activity; Franz et al., 2003 describes MALACH-related work.)

3.4.2.5 Information extraction

This encompasses attempts to identify relationships between entities in one or more documents (e.g. coreferencing) and the extraction of domain specific event types (e.g. free kicks, goals in football matches). The MUMIS project at Sheffield University is attempting to extract such information across multiple, multimodal sources. The problems of information extraction are discussed in Cowie & Wilks (2000) and Grishman (1997); the problems of information extraction from errorful automated speech recognition transcripts are considered in Grishman (1998).

3.4.2.6 Other

There are preliminary investigations into the extraction of language information (see e.g. the 2003 or 2005 NIST language recognition evaluations described in papers such as Martin & Przybocki (2003), dialect information (also addressed in the language recognition evaluations, with some sites treating each dialect in the same way they would treat a distinct language (Chen, 2006), emotional information (e.g. speaker state) (see e.g. the useful list of emotion related projects maintained by the EU HUMAINE project (Humaine, 2003-6), dialogue act information (see e.g., Wright (1999) and Webb et al. (2005)), and prosody.

Progress in the first four of these annotation processes has been driven by evaluations. Named entity, topic and information extraction techniques have been most heavily investigated for text rather than errorful transcriptions of unplanned speech – developing more robust techniques for handling such text is the subject of ongoing research (Ostendorf et al., 2005). A few category 1-2 named entity

extraction tools exist, but most of the rich metadata annotation research above falls into categories 4-5 and much of it has only been investigated for speech from a small set of domains, such as broadcast news and conversational speech. Emotion related work in particular is very preliminary and falls into category 5.

3.4.3 Music transcription

For years, scholars have anticipated a tool which could transcribe musical performances to music notation. Indeed, the original aim of one of the earliest and best known projects in musical computing was a system which would transcribe the performance of a musician on a specially designed music keyboard into music notation (Longuet-Higgins, 1976). A tool which automatically transcribes even a simple musical performance into correct and accurate music notation remains a distant goal, however. Perhaps this should be no surprise, since only highly trained musicians can make any such transcription at all, and even so the process involves a high degree of approximation and guess-work. On the other hand, transcription into some form of notation which gives useful information is possible for restricted kinds of musical sound, and it can be a useful tool in, for example, ethnomusicological research where systems like the melograph (a device which derives a continuous pitch curve from monophonic sound) have been in use for some time. A recent review of the state of the art in music transcription is (Klapuri, 2004).)

3.5 Analysis

The location of 'analysis' in the diagram indicates our intended meaning for the term: while many of the tasks and processes of annotation and transcription are in some sense analytical, we mean here that part of research where the results of annotation and transcription are subject to the judgement and intervention of the scholar who seeks to extract useful information, draw lessons, and form conclusions.

3.5.1 Analysis of audio and music

With respect to sound, the most significant contribution of ICT tools to analysis has been in the now almost routine extraction of frequency-domain information from sound signals. Analyses which focus on acoustic properties, in phonetics and music, regularly make use of tools which employ Fourier analysis or other methods such as auto-correlation to determine the component frequencies of a signal and their relative strengths. In the case of non-static sounds, this information is most commonly presented in a sonogram (a two-dimensional display with time as the horizontal axis and frequency as the vertical). Many such tools exist to effect such analysis: Wavesurfer (Sjölander & Beskow, 2006) is a good example of software from the research community, while Matlab (with its Signal Processing toolbox (The MathWorks, 1994-2006)) is probably the most commonly used commercial software. Musicians use such

tools for many purposes, including the analysis of instrumental tone (e.g., Fitzgerald, 2003) and the analysis of pitch articulations and vibrato in performance (Rapaport, 2004).

The analysis of musical performance has become a topic of considerable interest, spurred by the two factors of a now substantial history of recorded music and ICT tools to facilitate the analysis of music-as-sound. (The most distinctive project in the UK in this area is the Centre for the History and Analysis of Recorded Music (CHARM), at Royal Holloway and King's College, University of London.) However, no distinct set of ICT tools seems to meet the needs of researchers in this area. It would appear that there are still considerable gaps between the information which software can derive and present about musical sound and the information which researchers want to discover. For example, it is rarely a simple and straightforward matter to distinguish where notes begin and end in a sonogram, and while information on the precise frequency composition of a sound can be derived, that does not always correlate simply with its perceived pitch composition. The most effective use of ICT in this area, therefore, comes when software can automate some of a task or present information in a manner which allows the researcher to bring to play more effectively or more rapidly his or her musical ear and judgement.

A nice example of this is a simple piece of software, MATCH (Dixon & Widmer, 2005; Dixon, 2005) (category 2), which aligns two performances of the same piece, bringing two benefits. One is continuous data on the relative timings of the performances: a researcher can quickly discover if a longer overall performance results from a slower tempo throughout or from longer pauses at certain points, for example. The other is that the alignment facilitates simple switching from one recording to another at equivalent points in the piece. Thus a researcher can quickly and easily compare how two performers treat the same passage of the piece.

A second example is an equally simple suite of programs, to be used in conjunction with the music-analysis software Humdrum (Huron, 2002) and a sound editor, to capture information about the timing of beats in a performance, (Sapp, n.d.). Software to automatically recognise and track beats does exist (mentioned above under automatic annotation), but none has yet reached a level of accuracy and reliability where it has been adopted as a research tool. Researchers still wish to determine by ear exactly where the beats fall. To use the software, the researcher taps the beats as the music is played back. The time of each tap is recorded by the software, and can subsequently be adjusted in a sound editor in parallel with the original music. (The suite of programs is thus a form of manual annotation software.)

For music scholars, 'analysis' generally refers to the distinct sub-discipline of 'music analysis' which examines the structure of individual musical compositions in depth. This generally depends on the score as the primary source, and so does not fall within the remit of this report as dealing with audiovisual materials. There is no particular reason, however, why analyses which take musical sound as the primary source should seek to examine aspects of performance (as in the CHARM project, for example) rather than the details of specific pieces. Indeed, for popular music and electroacoustic music, there generally is no primary source other than the sound, so music analysis which examines the sound

is most appropriate. We can expect, therefore, that ICT tools, perhaps intended originally for MIR or for the analysis of performances, will come to be used in the sub-discipline of 'music analysis' also.

3.5.2 Analysis of film

Two main avenues of software-augmented analysis of film and video exist. The first seeks to automate analysis of the visual forms, and narrative structure of film and television. The second uses databases and presentation software (media players mainly) to facilitate new kinds of analysis. The two avenues have not yet converged. It will be interesting to see what happens if they do.

As for the first, tools for automated analysis of visual content, as mentioned above in the partitioning discussion, exist already. Some analysis is done in the interests of indexing and searching. For instance, the Virage VideoLogger software claims to automatically create structured indexes of content: 'At the same time video is being encoded, VideoLogger's advanced capture and analysis technology works in real time to automatically create a structured index about the content. Time-synchronised to every encoded copy made, the index enables immediate, accurate search and retrieval of assets. In addition, because the video is data-driven, it can then be tied to applications for revenue generation, enhanced collaboration and expedited communication' (Virage, 2006) (category 1). More ambitious projects try to provide a semantic analysis. The MoCA Project (Automatic Movie Content Analysis) (Praktische Informatik IV, 2006) (category 3) seeks to provide automatic identification of the genre of a film by comparing visual statistics of frames and sequences with genre statistical profiles.

To date, the main software technology used in analysis of film and television has been the database coupled with DVD. The 'Digital Hitchcock' project, by Stephen Mamber (UCLA, n.d.), represents a well-known early example. It represents all 1,100 shots in the Hitchcock's *The Birds* alongside Hitchcock's storyboard illustrations. Using commercial multimedia authoring software such as Macromedia Director, various projects such as the Labyrinth Project (Kinder, n.d.) have used a combination of presentation technologies (Quicktime, Flash, Shockwave) and online databases to analyse narrative in feature and documentary film.

3.6 Presentation

At almost every stage of the research process, researchers make use of different ways of visualising, summarising or tabulating audiovisual materials. *Presentation* refers to all the different ways in which digital technologies display or render different audiovisual materials apart from simply reproducing them. For instance, the timeline in a video editor or the waveform in a sound editor are presentations of images and sound respectively. In this sense, a Microsoft Powerpoint show is not really a presentation as such. Presentation is closely linked to analysis. In some ways, we could say analysis is nothing but a process of generating increasingly complex, conceptually ordered presentations.

3.6.1 Summarisation

It is not always appropriate to play back a full recording or clip to a user. (Consider filtering texts by eye versus filtering a set of audio clips: the latter is usually more time-consuming and not necessarily appropriate in a search system.) There has been some, albeit limited, work on generating summaries or shorter but informative representations of spoken word content, mostly in category 5. The limited amount of research may be due in part to difficulties in evaluating the quality of summaries, since the desired properties of the summary will often be application- or even user-specific.

A summary may be generated in one of three ways: using the audio alone, using an automatically generated speech-to-text transcription alone or using both the audio and the speech-to-text transcription in combination. Techniques using the audio alone include time compression techniques such as eliminating silence or speeding up the clip (often maintaining pitch for intelligibility) (e.g., Tucker & Whittaker 2005): the resulting compressed signal can then be played back. Techniques operating from errorful transcripts may be direct adoptions of techniques for general text summarisation (as is apparently the case in Pickering et al., 2003), though performance might be improved by incorporating knowledge of the errorful input (Furui, 2005b).

Techniques exploiting both audio and transcript include the work by Koumpis and colleagues, who use both lexical and audio derived prosodic information to identify elements to include in the summary (Koumpis & Renals, 2001). Summaries generated from transcript or from both audio and transcript can be presented in textual form or in audio form (perhaps involving the use of the speech synthesis system). The choice of summary presentation format is likely to be application dependent: as an extreme example, audio summaries might be useful in situations such as over-the-phone access to a voicemail box, whereas textual summaries could be sent to a mobile phone via SMS (as in the VoiSum system (Koumpis & Renals, 2005)). Spoken word content summarisation and usability issues have been considered in some detail by Arons (1997) and by Furui (2005b).

The same considerations have motivated research into automatic summarisation of music. The common approach is to perform some kind of self-similarity analysis of the audio signal, often by means of a frequency-domain transformation, and then to extract those segments which are similar to other segments. These are likely to correspond to recurring passages such as the chorus of a song, and so to contain music which is salient and typical of the whole. A short segment of audio can then be constructed by stringing together characteristic extracts (see, for example, Peeters et al., 2002). There are many complications and issues in the process, however, and no tool has yet advanced beyond category 3.

Summarisation is not always necessary. For some applications such as the display of search results, users may tolerate presentation of short search-matching sections of an errorful transcript together with an audio playback option for verification (described as ‘What You See Is Almost What You Hear’ interfaces (Koumpis & Renals, 2005)).

3.6.2 Speech-to-Speech Translation

In systems somehow supporting cross-language searching, the issue arises of how to present material in some foreign language to a user who queried in a different language. Aside from the most obvious solution of involving a human in this process, a recent line of research ('speech-to-speech translation') has addressed the translation of spoken content from one language to another. To date, this work has mostly addressed constrained domains far from that found in a typical audiovisual archive (such as travel, emergency medical diagnosis, defence-oriented force protection and security. The IBM MASTOR 'Multilingual Automatic Speech to Speech Translator' project (IBM, n.d.), for example, falls into category 2. Speech-to-text translation systems (see Sections 3.4.1 & 3.4.2) are also only in the category 4-5 stages of development. However, developments in these areas may be relevant to future digital libraries projects.

3.6.3 Visualisation

Often it is useful to present the information in or derived from audiovisual material in some other graphic form either to enable overall patterns or structure to be seen, or to assist in the identification of points of particular interest. The topic is particularly common in music research, where systems which enable one to 'see' music of which our experience is otherwise ephemeral. Discussion of different kinds of music visualisation are given in Isaacson (2005). The efficacy of visualisation is demonstrated in a study which examined the degree to which providing a visualisation of various acoustic properties ('spectral magnitude', 'novelty', 'rhythm magnitude') aided users in finding particular points in a music recording (Wood & O'Keefe, 2005). The study concluded that visualisations did indeed help in navigation, but visualisation of different properties was most effective for different pieces of music.

Obviously, the kinds of visualisation and their level of detail will vary from one purpose to another – the whole point is to present the data in a form suitable for the particular project – but some commonalities do emerge. One is the repurposing of editing software to produce visualisations of the composition or structure of some material. Film scholars use commercial software such as FinalCutPro and AdobePremiere not only to edit digital video footage (for example, to extract clips for presentation or personal archives), but also as a way of examining the composition of film at various levels. The editing timeline is a central component in most video editing software, representing the complete set of frames in a film. Using the timeline, scholars can zoom in and out from frames to the overall film, and also view overall structure of the film or analyse transitions between shots.

For music, scholars similarly use audio editors for visualisation. Every audio editor shows a 'waveform' display of the signal, showing the peaks of the sound pressure wave or (at higher resolutions) the actual wave itself. This provides a quick and easy way of spotting sound and silence, and sometimes allows the beginnings and endings of sound events to be found also. It is not uncommon for audio software to generate sonograms also, which show the power of different frequencies in a two-di-

mensional display, often using colour, but the degree of fine control over their generation varies, as does the ability to export the resulting data.

Specialised software involving visualisation exists also. Video editing and mixing tools developed for Vjaying (selecting and mixing found video materials, and setting them to music) have addressed the problem of how to rapidly select and organise quite large collections of film and television footage. Software such as Resolume (n.d.), an instrument for live video performances (category 1), allows rapid selection, changing, combining and comparing of video clips on screen. For music, there are examples of projects which attempt to show higher-level or more 'semantic' qualities in the audio stream. Examples are provided by aspects of the CLAM Music Annotator, mentioned above, which includes panels to visualise automatically extracted data on harmony and tonality in a time-varying two-dimensional colour display where neighbouring regions are associated with related keys or harmonies and colours indicate strength of reference (Gomez and Bonada, 2005). A similar tool, which explicitly allows the development of new 'plug-ins' for new visualisations, is the Sonic Visualiser from Queen Mary, University of London (Centre for Digital Music, n.d.)

Visualisation has also been used to show relations between pieces of music rather than within them. In particular, a number of projects use it as a means to solve the apparent problem of how to navigate a very large collection of popular-music recordings, related by various similarity measures, using metaphors of contour maps (Mörchen, Ultsch, Nöcker & Stamm, 2005), or factors like size and colour (Goto & Goto, 2005). In this, however, they do not significantly differ from data-visualisation projects in general.

3.7 Integration

It should be clear from the above that searching, annotating, transcribing, analysis and presentation are not discrete, atomic operations. However, very few attempts have been made to develop technologies that combine all aspects of research. Again, tools designed for working with speech lead the way. Apart from a few large scale research-led development projects, technology that integrates different aspects of the research process does not yet exist for working with music and video. The large-scale integrated projects which do exist are broad in scope and different aspects of the projects typically fall into different development categories.

3.7.1 Malach (Multilingual Access to Large Archives)

The Survivors of The Shoah Visual History Foundation (VHF) was founded by Steven Spielberg to enable survivors of the Holocaust to tell their stories and have them saved in a collection that could be used to teach about intolerance. Over 52,000 testimonies (116,000 hours of video) have been collected, containing 30 languages and forming a 180 TB digital library of MPEG 1 video (Gustman et al., 2002). This has been manually catalogued, although not in the detail that was originally planned: the very de-

tailed and extensive human cataloguing that had been envisioned and was completed for over 3000 testimonies was found to take about 15 hours per one hour of video, which even with good tools was estimated to cost over $150 million. The foundation therefore backed off to a 'real-time' cataloguing methodology, in which one minute clips are linked with descriptions and person objects. In parallel with this, a significant research project (the MALACH project, n.d.) investigated methods for fully or partially automating cataloguing. The project addressed issues such as automatic speech recognition to automatically generate transcripts for speech in multiple languages, automatic segmentation of transcripts into shorter units suitable for retrieval, automatic classification in order to assign thesaurus terms to segments, automatic translation in order to allow querying in multiple languages and also undertook user studies in order to investigate the types of user access that would be most useful to users of the collection. This represents a very challenging domain for speech recognition, since the interviews contain natural speech filled with age-related coarticulation due to the age of the speakers, heavily accented language, and uncontrolled speaker and language switching by often very emotional speakers. The automatically generated metadata was perceived to be usable by the project investigators, providing a flexible means of access for users whose needs may not have been addressed by the metadata schema used in manual annotation (Byrne, 2006). One of the major contributions of this project was an information retrieval test collection for spontaneous, conversational speech (625 hours of automatically transcribed speech) based upon real information needs derived from user requests. This should provide a standard test for conversational speech indexing systems. (Oard et al., 2004).

3.7.2 Variations2

The Variations2 project, from Indiana University (2005) was initiated as research to establish 'a next-generation [system] for research in digital library system architecture, metadata, network services, usability, intellectual property rights, and music pedagogy.' (Dunn et al., 2006) The Variations2 system, as deployed, allows access to recordings as well as scanned and encoded musical scores. This clearly ambitious, end-to-end system is grounded by a mature, tested metadata model, in which *contributors* (performers) create *instantiations* of *works* (which, in turn, are by composer *contributors*). *Instantiations* (performances) appear on a *container* (such as a CD), which may give rise to other *media objects* (an encoded MPEG file). This model is as simple as the world allows it to be, but is powerful enough to enable such further developments as searching, collaborative annotation, and automated and manual analysis of music. These and other rich metadata definitions have been designed to be explicitly suitable for classical music, are much more complex than the categories of the Gracenote database, and grounded in information-science research.

The Variations2 tools are mostly based on cross-platform Java applications. The audio tools include a timeline tool to support formal analysis of works, allowing users to add timepoints (annotations) between sections, which can be visualised as bubble-like arcs, which are familiar to those studying musical structure. Although the goals tend more towards pedagogy than research, Variations2 is

one of the best examples of an end-to-end music information retrieval system with scholarly underpinnings.

3.7.3 Informedia Digital Video Library project

The Informedia Digital Video Library (categories 2-5) (Carnegie Mellon University, 1994-2006) represents one of the most ambitious projects in the space. Funded by both the first and second phase of the NSF Digital Library Initiative, it has the overarching goal of achieving automated machine understanding of video, including search, retrieval, visualisation and summarisation in both contemporaneous and archival content collections. Informedia has aggregated a library of multiple terabytes of video, mostly broadcast television news and documentary content. The first phase of the project integrated technology for speech, image and natural language understanding to automatically transcribe, segment and index broadcast video for searching and retrieval purposes as seen in the news on demand application which automatically processed broadcast news shows for the Archive. The second project phase investigated techniques and video information summarisation and visualisation, extending single video abstractions to summarising multiple documents in different collections in visualising very large video data sets. Separate projects investigated a variety of tasks including multilingual broadcast news archive and cross-cultural archives, including the connection to the ECHO project in Europe and collaboration with the Chinese University of Hong Kong. Related work is ongoing.

Informedia is notable for its wide-ranging exploration of the space, but also (as noted in a recent paper about lessons learned during the 10 years of the project) it has derived an infrastructure that allows daily processing without any manual intervention. This distinguishes the system from many of those described elsewhere, which are often research-deployed only or deployed with a limited number of users. Informedia has developed a robust grouping of components and experienced problems that do not arise when investigating single research issues including identifying techniques which are too computationally expensive, those which are overtrained to a particular dataset and those that go out of date over time (Hauptmann, 2005).

Table 1. Uses of audiovisual media within the arts and humanities

	Research resource	*Work record*	*Research outputs and/or dissemination*	*Teaching and Other*
Self recorded or constructed	E.g., linguistics corpus (1), oral history interviews (2), auditory archaeology recordings (3)	E.g., archaeological excavation recordings (7), raw anthropological/documentary footage (8)	E.g., multimedia archives created for use by researchers (9), technical/scholarly/popular presentations of research results involving multimedia (10,11)	E.g., phonetics sound examples for class, tutorial exemplars of form
Found	E.g., linguistics corpus (4), films/television/radio for historical or cultural analysis (5), poetry readings (6)			E.g., clip examples for teaching, examples illustrating responses to media questions, contributing to external projects

3.7.4 National Gallery of the Spoken Word

Similar issues are being investigated in the NSF digital library initiative project developing the National Gallery of the Spoken Word (n.d.). The project as a whole is investigating issues related to digital watermarking, digitising and categorising, copyright, distribution and educational program development for 60,000 hours of historical recordings from the 20th century. One specific part of this problem is investigating the recognition of and search within this data and has developed an experimental online spoken document retrieval system called Speech Find. The challenges of this collection include the variety of recording technologies, acoustic environments, speaking styles, names and places, accents and languages and the time varying grammar and word usage (Zhou & Hansen, 2002; Hansen et al., 2001).

4 Appendix C. Researchers: practices, possibilities and expectations

The preceding technology survey described technology capabilities. This section considers the ways in which the needs of arts and humanities researchers might be satisfied by some of the technologies surveyed. The section begins with a brief snapshot of the ways in which audiovisual media are currently used within the arts and humanities. It then presents the results of our qualitative study of researcher needs, using scenarios to demonstrate how some of the technologies discussed earlier might meet those needs.

4.1 Snapshot of Current Humanities Uses of Audiovisual Media

Table 1 illustrates some of the ways that audiovisual media are currently being used within the arts and humanities. (The numbering relates to the examples that follow.) The following examples may provide useful illustration: [n.b. the examples were selected to show the range of possibilities and do not reflect any assessment of their academic merit]

1. Linguistics corpus: the IViE ('Intonational Variation in English') project investigated cross-varietal and stylistic variation in English intonation using self-recorded data from nine urban dialects of British English from both male and female speakers (Grabe, 2003).
2. Oral history interviews: the 'Childhood in Russia 1890-1991: a Social and Cultural History' Project (Kelly et al., 2004) has collected a variety of detailed, tape-recorded interviews with informants from a range of different generations.
3. Auditory archaeology: There is an experimental auditory archaeology (Witmore, 2005) project at the Catalhoyuk site. Volunteers engage in activities such as sweeping, polishing walls, making plaster, applying plaster and repairing a platform within an experimental house and do so using materials and techniques supported by Neolithic-era evidence. They wear binaural microphones to record the sounds reaching their ears, creating stereo digital recordings which are transferred to a PC setup for storage and analysis (Mills, 2004).
4. Linguistics corpus: the British National corpus is an approximately 100 million word corpus of present-day British English from a variety of sources, containing both a spoken and written component (University of Oxford, 2005). It has acted as a found resource in a number of studies, such as a study investigating variation in vocabulary usage according to gender, age and social group (Rayson, 1997).
5. Films/television/radio: uses of such media as a found resource include a project comparing British and German film propaganda during the Second World War (Fox, 2003). Popular and/or high culture as found on commercial CDs and DVDs form a large part of the primary materials for numerous researchers.

6. Poetry performances: the study of poetry readings (and of silent readings of poems) is discussed in 'How to Read a Poetry Reading: Reading the Reading' (Middleton, 2003). This paper is linked to the British Electronic Poetry Centre (BEPC, 2004) which provides information on poets, their publications and some audio files demonstrating their work (a potential found resource for future research).
7. Archaeological excavation recordings: the long-term Poggio Imperiale project was associated with the creation of an archaeological and monumental park investigating a hilltop to the west of the Italian town Poggibonsi. Digging was filmed with a video camera and the resulting recordings were assembled and edited using desktop computers (e.g. with QuickTime software) (Archaeological Computing 1996). Another excavation, at Catalhoyuk in Turkey, also has a project investigating video recording within archaeology (Cee, 1996).
8. Some audiovisual researchers are practitioners, creating new research in an electronic medium. Their raw, intermediate results represent a form of work record from which new products are distilled.
9. Documentary footage: the Designing Shakespeare audiovisual database (AHDS 2005) includes a collection of researcher-recorded video interviews with designers, in addition to a text database of production details and theatre review excerpts, a collection of production photographs and VRML theatre models.
10. Archaeology walkthroughs: the use of visualisation tools such as pre-computed video walkthroughs is discussed in the technology paper 'An Interactive Photo-Realistic Visualisation System for Archaeological Sites' (Chalmers, 1996).
11. Scholarly publications including audiovisual components: the Sphakia Survey was an interdisciplinary archaeological project which attempted to reconstruct the sequence of human activity in a remote part of Crete (Greece) between 3000 BC and AD 1900. The project made a 50 minute video about the Survey (Nixon & Price, 2000). Although primarily for use in university classes, it was also used to report to general audiences by through national television networks in Greece and elsewhere and through distribution of individual copies (Nixon & Price, 2004).
12. Communications for public consumption: academics often contribute to broadcast media productions such as The British Empire in Colour (BFI 2002), as seen in an interview with the production team (Luscombe, 2002).

4.2 User Needs Study

The goal of the user needs study was to determine ways in which new and emerging tools for time-based AV analysis, annotation and search might aid humanities researchers, either by facilitating their exploration of conventional research questions or by enabling them to ask new research questions.

4.2.1 Methodology

Interviews with academic researchers in the humanities were conducted in three phases from October 2005 to July 2006:

Phase 1 aimed to interview one person per humanities field, using the AHRC Research Subject Coverage for guidance (AHRC 2003). Phase 1 interviews were loosely structured using an interview questionnaire, supported by PowerPoint props demonstrating screenshots of the following tools uncovered in the early phases of the project:

1. BLINKX: a live system supporting browsing and free text search for AV on the web (Blinkx 2006)
2. ANSES: a demo interface for news summarisation including automatically extracted organisations, people, locations and dates (Pickering, 2006)
3. FERRET: a meeting browser tool (IDIAP, 2006)
4. MULTIMODAL ANNOTATION TOOL: a manual annotation tool for video including associated soundtrack (Adams 2002)

The initial interview questions explored a researcher's current usage of time-based AV and any difficulties they experience when working with AV. The interview then moved to a short demonstration of tooling possibilities using the PowerPoint props. The tools shown were selected to stimulate exploratory discussion about user needs such as access to AV on the Web or in archives (BLINKX, ANSES), non-linear access to AV (FERRET) and annotation of AV data (MULTIMODAL ANNOTATION TOOL). The tools demonstrated were chosen because of their online accessibility, rather than any criterion reflecting technical merit or humanities-specific design; this meant a handout containing the appropriate links could be distributed at the end of the interview enabling an interested researcher to explore further if they wished. Discussion was not limited to these classes of tools and often led to quite unexpected tooling suggestions that reflected the needs of individual researchers.

Phase 2 aimed to interview modern historians whose web presence suggested audiovisual data might be a potential resource (even if not currently used). Phase 2 interviews were aimed at gaining a more detailed understanding of the work process of researchers in one specific field, extracting information about their use of resources by asking researchers to talk through a typical research project prior to the exploratory discussion of the PowerPoint props. This information clarified the ways in which some of the tools under consideration might fit into the research process.

Phase 3 interviews concentrated on researchers whose primary interest was in audiovisual material, such as films in popular culture, music within films, video games, or general musicology. The interviews were conducted using a set repertoire of questions on audiovisual media usage and research practice, garnered from experience with the first two phases. Presentation of technological tools was based on knowledge of the literature, and was presented orally as called for in the situation.

Additional interviews were conducted with the BUFVC (British Universities Film and Video Council), a creative video artist, a technologist collaborator of a music researcher and a Modern Languages IT Manager. These interviews provided useful background for interpreting the main interview results.

4.2.2 Institutions represented

Interviewees came from the RSAMD, Glasgow School of Art, Royal Holloway and Goldsmiths, University of London, and the Universities of Oxford, Reading, York, Sheffield, Manchester, Edinburgh, Glasgow, Kent and Lancaster.

4.2.3 Subjects represented

Phase 1 interviewees were drawn from a wide variety of humanities fields. The creative arts and art history, law and philosophy were not represented. Phase 2 interviewees were all drawn from modern history. Phase 3 concentrated on researchers whose work focussed on music- and video-as-artefact.

4.2.4 Limitations of study

The study is small-scale and qualitative, with no verification of results. It covers only a small sample of academics from a small number of institutions, so cannot be assumed to be representative of the whole UK humanities community. There is an additional bias towards lecturers and professors, rather than graduates and research fellows, which may be reflected in the current research approaches and ICT uses that are discussed in the scenarios.

The use of canned (pre-stored) screenshots rather than live demos as interview props allowed interviews to be kept within an average of one hour, which was felt to be the maximum that could reasonably be asked of researchers who already have heavy workloads and are being asked to participate voluntarily. Their comments therefore reflect their interpretation of the canned demos rather than any practical assessment of the tools. Specific misconceptions that arose are discussed in the Section 4.4, 'Technical Expectations'. The props used also demonstrate general-purpose tools rather than those designed for humanities research purposes or operating on humanities-relevant data. This required interviewees to extrapolate to their own situations: however, on balance, this approach appeared to be quite effective, perhaps more so than when placed in a position of 'selling' humanities-specific tools.

There were occasional communication difficulties during the interviews due to differences in language usage by humanities researchers and the engineering-educated interviewers. Any errors of interpretation are the responsibility of the interviewers.

4.3 Interview Results

The combined results of all the interviews have been used to generate the following quasi-fictional scenarios. These summarise uses of analysis, search and annotation tools that were suggested but also mention some of the associated challenges to deployment (technological or otherwise).

We use the following notation:

[1] quote extracted from interview transcription;

[2] minor rearrangement or modification of the transcribed words of an interviewee for clarity, brevity and/or to make anonymous by substituting for one or more [identifying phrases];

[3] paraphrase constructed from handwritten interview notes;

[4] quote extracted from email exchange;

Composite scenarios are constructed from multiple interviews and/or e-mail exchanges.

4.3.1 Obtaining research resources

4.3.1.1 Self Recorded

Composite Scenario: Researchers X, Y and Z all have sets of cassette recordings sitting on the shelves of their offices, from earlier work collecting oral history interviews, ethnographic interviews and data collected for a sociolinguistic study. They would like to convert them into digital form to avoid problems of further cassette degeneration and so that they are more easily accessible for their own reuse; X would also like to send his interviews to transcribers in digital form and ultimately to put his collection online (after removing any confidential sections or interviews that the participants do not wish to be made public) in order to encourage reuse by the wider research community. They would like help or instruction about how this should be done.

Scenario: X is a documentary filmmaker, and, during a preparatory visit to the documentary site, brought a DV camera along on a visit to the tourist officer for the town he was profiling. 'There's the very important interview that I probably will end up using… I was just going to talk for an hour to the head of tourism… [I thought,] "Oh, I'll take my camera along," [and I asked] "Do you mind if I shoot?" "Oh, go ahead," and so then I shot it, and then I had it on video, which I wouldn't have had otherwise, and now it becomes a different sort of resource. That's pretty amazing that you can do that these days, with smallish equipment so it's not intimidating to people… So technology, the way that it's developed, has worked much more closely with my own methodologies, interests, and the way that I like to film. It's made it a heck of a lot easier.' [2]

4.3.1.2 Found Data

4.3.1.2.1 Online AV and Web AV Search Tools

Scenario: X is writing a textbook about a diaspora community that has spread around the world, including into the UK. He has a well-established, theme-organised index card system for recording *information* about useful sources he has located, whether in archives in the UK or abroad or on the Web. He is familiar with search engines for text: 'One of the interesting things about Google is that you've no idea what you'll find' [1] He's recently come across some of the new search engines for online audio and video, and they will fit readily into his research process with a low learning curve since the interface is similar to those he has seen for text-based web search engines. Since they index contemporary news information and podcast information, they are immediately useful for his textbook project: a brief search turns up newscasts about the role played by community members in recent factory strikes and a series of podcasts from individuals discussing community issues. He wonders though about how audi-

ovisual data might fit into the dissemination process: 'If I found relevant video, I'd then have to transcribe what they say onto an old-fashioned medium: I don't think many academic publishers have the notion of linking books to multimedia, so you'd be working with transcripts.' [2] However, he notes that 'What one can find is only as good as what is put in' [1] on the Web. Also, the audiovisual search engines won't be useful to some of the other research questions he's exploring, since these relate to prominent historical figures captured primarily in text media and photographic images: '[time-based] audio and video are not a particularly useful resource for this because the other sources of evidence are very strong and deal with the things I'm really interested in.' [2]

Scenario: X is a modern historian exploring the responses of individuals from other cultures to the events of 11 September 2001. He began his research by exploring text-based blogs but has now realised that search engines for audio and video blogs could provide him with additional sources.

Scenario: X researches the modern history and politics of an eastern country. His projects are both 'historically-based in the classic sense but also about the impact of history upon the contemporary [country being studied].' [2] The historical part of his work tends to involve 'things which are thought of as standard for historians, e.g. going to archives in various countries particularly in [the country] but not exclusively in [the country] to find materials which are stored there, unpublished materials as well as using things that are available in the collections and libraries around the world. I also go to places to do interviews with some of the people who are involved in contemporary politics … and also use resource bases such as newspaper holdings and occasionally some contemporary cultural uses of history such as films, lectures, TV shows which are often on DVD. The majority is published or unpublished written sources.' [2]

He sees the potential in video search engines (already available for the country he is interested in, in the major language) in looking for news broadcasts and contemporary recordings from his desktop. However, there are other types of AV material he would find useful. 'If there was a sort of historical searchability or old newsreel footage or whatever put up on the Web, that would certainly be very useful. The contemporary stuff would be more than enough for my contemporary work but not for a historical project. One of the things I am looking at for my new book is the way in which propaganda was shaped in [the country] during wartime. The vast majority of the material I am using will end up being mostly if not all newspaper and print stuff because that's what survives and is relatively easy to get hold of in archives and so forth; if there was somewhere you could actually look at say wartime cinema propaganda broadcasts … or radio broadcasts then that would be a really useful addition to that. At the moment I'm not really aware of any user-friendly way in which that could be done: there probably are stores and stores of this sort of thing in some archive in [the country] but the effort … unless it was really your specific subject that you really wanted to push … I think most people would think that life is short and you wouldn't make that extra effort whereas if I was able to call it up as a resource in the way you can do a digitised newspaper or something of that sort then obviously it would be more attractive.' [2]

Scenario: X has a theology background. 'Audio and video isn't used as a matter of course, but certain areas may use it. For example, historical theology has links with archaeology and may use simulation and video modelling techniques. It may also have a place in sociological approaches to theology, perhaps contemporary theology or pastoral theology.' [1] Its use remains unusual in his opinion. 'However, it has potential application. In contemporary theology, so much primary research material is generated through the popular media and popular culture. Or the studies of how liturgy and how theology happens, studies of how it happens in particular contexts. Recordings of liturgical events as ritual.' [2] The advent of AV search tools combined with web-based collections that can be accessed from the researcher's desktop may prompt new research: 'The average theologian will not be able to think of any collection that could be indexed; however, he may be able to think of lots of uses for someone else's collection. It's the combination of search tools plus collection that sparks ideas. New kinds of access to an existing collection in digital form could make it yet another primary resource useful for research e.g. to explore questions relating to new religious movements or sociology of religion, contemporary more social science aspects of theology. This technology could be a way of encouraging reuse (perhaps by licensing) of collections. The comments that this might be useful are very much based on the assumption that there would be online, desktop access to these collections though.' [2] He concedes that 'the application to work in fields such as early Christianity is not quite obvious.' [1]

Scenario: X is analysing the culture and history of a non-European region. He doesn't have easy access to the region's time-based media such as films or television from his base in the UK. 'Archives in [the region] are not easily accessible to outsiders and, quite apart from secrecy, I'm not sure how much audiovisual data is well preserved.' [3] He has accumulated his own collection of films by purchasing them when on holiday in the area (usually on video cassette) and stores these on shelves in his office. He typically digitises these films using his own hardware and, where necessary, subtitles them into English and/or extracts clips of interest for research or teaching using a tool such as iMovie. He also uses a variety of other resources, including 'periodicals, articles on the Web and so on.' [3]

The research questions he addresses are varied. In some pieces of research he analyses known films as a whole e.g. understanding the narratives. The set of films he could consider would be expanded if and when the region's audiovisual outputs become available online, whether for purchase, streaming or some other access mechanism. In other pieces of research he addresses questions which may be answered using a variety of sources of evidence. For example, his current research project is considering issues associated with the commercialisation of religion and may draw upon sources including religious imagery, advertisements and televised discussions. He immediately sees the potential for audio and video search engines for helping the latter kind of project, enabling a more efficient search for relevant clips on the Web e.g. through queries suggestive of 'advertisements, religious programming and reality TV, televised discussions about consumption,' [3] and he would be willing to be creative in coming up with queries in order to find useful data. However, this potential will only be realised when sources from his region of interest are put online and when search engines become available for data in the corresponding language.

Scenario: X is interested in discourse differences across a number of non-Western countries and is currently exploring issues relating to visual grammar and reception. His research begins with a process of dataset construction, which requires him to locate sources of moving image data and then filter that data in order to find instances of desired 'events' in the soundtrack or leading imagery, such as clips showing a weapon or alluding to a weapon. These instances form the dataset for his research. At present, he primarily obtains data through off-air recordings (e.g. made by colleagues in the region) or from the few academic, area-specific websites online: more online sources of data from the area would benefit his research, particularly if easily locatable through search tools. The filtering process is currently very time-consuming, requiring a full viewing: search tools that could help him identify relevant 'events' within videos would be very helpful in speeding up this filtering process. He notes that the search would not need to operate perfectly because although he needs to find a number of 'events' it is not essential to discover all of them. The envisaged search tools could be web-based and allow him to search within AV on the web, combining the location and filtering stages of dataset construction; an alternative would be a package that could index his already downloaded data collection and speed the filtering stage. In either case, the tools would need to support search in his language of interest and perhaps an image-related search as well as a free text search. Since these envisaged tools do not currently exist in a packaged form, he sees manual tools as a potentially useful and available alternative for the filtering step: a tool such as the IBM annotation tool could support the process of marking up and categorising soundtrack segments or image regions and this could be combined with a viewer tool which supports the recall of items in the same category (e.g. the category of clips showing a weapon). Such viewers could be straightforwardly developed for certain formats, rights permitting. The researcher might also investigate annotation tools such as MediaMatrix (MATRIX 2005) and similar tools being developed by the social science community.

Scenario: X is interested in issues involving a contemporary composer and in the reception of 20th-century music. Although he doesn't currently make extensive use of spoken word audio or video, the new search engines for contemporary news and entertainment radio and television on the Web may offer access to relevant interview and performance review data from his desk.

4.3.1.2.2 AV Archive Browsing and Search

Scenario: X is interested in studying different performances of a play which are stored in an audiovisual archive. The archive is experimenting with a new display facility that displays random clips from the collection when a researcher works through the catalogue. The researcher happens to spot that some of these random clips show audience-related, rather than performance-related, information. This serendipitous discovery wouldn't have occurred with their traditional text catalogue and leads him to investigate audience changes in theatre performances over time.

Scenario: X is a modern historian investigating the social history of an English-speaking country outside the UK. He mainly uses traditional archives, but sometime uses tools such as Google Image Search to locate images that bring events to life for students. He doesn't currently make much use of time-based media. Very occasionally he'll go to archives and read their transcripts of potentially relevant

video and he has independently accumulated a few documentaries on prominent political figures in that country, but he finds it 'takes a lot of time to get a little way with video. It takes time to find the video, particularly when the collection is not online and I have to travel abroad to go to the archive, and it takes time to find what's relevant … It's much easier to scan text to find relevant sections than it is for video without transcripts' [3] He has analysed some propaganda videos in the past, though, and certainly sees uses for audiovisual data in the future if it became more accessible.

The ability to do a free text search within a single archive collection of spoken word data might encourage him to use collections which have not been transcribed, particularly if results are cued up around the query terms and linear scanning of full tapes is not required. Even better would be to search for archived or other audiovisual research data via the web from his UK office, particularly where systems return clips which are cued up to the relevant point, but he observes that many of the current web AV search systems do not index the kind of data that he needs for his research. Because he investigates mid-century social history from the 'bottom-up', he would be interested in recordings from that era involving 'the people' e.g. 'speeches by the regional mayor or activists … or the unedited raw footage collected by production companies might be useful.' [3] If this kind of data became more accessible, it would not just provide an additional source of evidence in answering existing research questions; he envisages addressing new questions such as televisual representation or comparative studies of representations of things in television, text and pictures. The latter would be interesting because, in his experience, most research in his area today cites national newspapers rather than national television stations, even though more people watch the latter and it is arguably more influential.

Scenario: X is a film and television researcher. He typically makes use of a number of resources. These include written sources such as encyclopaedias and so on: 'These are still key to research in this area since they are widely available.' [3] He also makes use of many other resources, such as his extensive department library of VHS recordings of films and television, digitised resources such as databases about TV programming and scheduling and numerous Web-based sources of information on film and TV (academic and otherwise).

For some projects, he starts with a constrained dataset – such as the works of the particular scriptwriter – which he watches carefully in order to construct a thesis and then revisits in order to find evidence to support or rework that thesis. However, there are other situations where he has broader information requirements. For example, he is interested in exploring the influence of another country's television programmes upon UK television. The archives of UK broadcasters contain useful resources for this project, but these are not currently accessible by everyone. Fortunately, he's able to make use of his contacts to gain access to one such archive and can use their very detailed internal catalogue in order to find television programmes covering relevant topics. 'Finding data through archives, though, is an art and having an inside contact that can help is valuable academic currency. Such contacts are not available to everyone at all archives and so facilities for archive search are potentially powerful for improving access to some collections.' [3] He also notes his work would be greatly facilitated by a UK copy-

right deposit requirement for audiovisual data, as exists for books: without this, research is not able to address the full breadth of data which is produced, only that which is accessible.

4.3.1.2.3 Commercially/professionally available AV

For a certain class of researcher, there is a great deal of stock placed in one's personal, commercially-obtained collection of CDs or DVDs. The stereotypical researcher in this class is concerned with popular culture and the audiovisual material in itself, but is not limited to it. There is a lot of focus placed on an individual researcher's collection, and it tends to be made up of personal purchases, enabling the resource to follow the researcher from institution to institution. There is a secondary interest in broadcast media, both as a source and as an inspiration.

Scenario: X is a researcher in a music department, focussing upon film music in contemporary 'art' cinema. She investigates films from the mid-20th century to films of a couple years ago. A film in particular from the 1950s has provided a certain degree of frustration, both in terms of the detail scholarship required to track down sources, and with the 'rewriting of history' on the part of the film studio in having released alternately uncensored and censored versions of the original film. Different versions are available at different times, and new releases can completely supplant older versions. The sonic and timing differences between PAL and NTSC releases (4% in length, or a semitone in pitch) is considered an occupational hazard in her field.

Scenario: X is a composer, researcher, and theorist in a music department, who, drawing upon theories of *play* and the *immediate erotic*, often allows himself to be seduced by a new piece or a new hearing through broadcast media. 'It's the chance meeting, it's the glance that catches your eye, that turns your head. So the radio's very powerful, and I owe [BBC] Radio 3 an enormous amount for setting extraordinary things before me…. For instance, I heard [a composition by a British composer]; [he] never set German very often, and never set Goethe, with one exception… there it was, [and my reaction was] "This is exactly what I need to hear, right now."' [2] Although the source is a commercial recording, there was never an explicit search for what ended up being an important piece for his research and writing.

Composite Scenario: X, Y and Z are all contemporary culture researchers who find some proportion of their material as video on DVD. The limited DRM system that governs DVD (the content scramble system, CSS) and theoretically prevents an unlicensed viewing device from *viewing* copies (but not making them) has been defeated years ago, but due to legislation throughout the western world, disseminating information about such things is illegal. Although all of the researchers had legitimate reasons for wanting to view transcoded video from DVDs, they were stymied from doing so by the fact that tools for doing so were driven underground, out of the mainstream. Much was made of the fact that these technological barriers now exist where there were none before. 'When I ask the AV services to edit an extract from a DVD for use in the classroom, they always ask, "Do you have a videocasette?" because, even though they have the technical know-how, *they* find it a headache to deal with.' [3] Future digital distribution formats promise more than headaches, but real barriers to normal scholarly use.

4.3.2 Data preparation

Composite Scenario: X is a historian. 'I don't use time-based media much at present … when I do, I use a notebook to generate a rough transcript such as '2 minutes: event Z happens' and then I can fast forward around if I need to revisit sections. If Web search tools for audiovisual data become available and the data I am interested in is online and exposed to these tools then I might use manual annotation tools on that data – to mark points and be able to jump back into those points could be helpful. But the search tools and relevant online data are key; the other tools just make things faster when working with AV' [3] Researcher Y is a film and television researcher who has similar needs. 'It would be useful to have a facility which would let one mark and categorise short clips for easy revisiting or for exporting e.g. for playing in lectures.' [3] A manual annotation tool with an adjustable lexicon of annotation categories combined with a viewing tool for revisiting annotated categories (as discussed in Section 3.3.2) could fulfil the bookmarking/revisiting function for some data formats; clip extraction is supported by existing tools for some formats though the right to do so may need investigation.

Composite Scenario: X and Y use popular culture as their subjects, whether in audio or video. After an initial, impression-gathering viewing of the material, both perform a rough timeline of the interesting and notable elements in the audiovisual material. There is not always a timecode-accurate notation of the start time, and it is rare that a shot-by-shot or event-by-event notation is made, but there is a fine level of granularity. 'I will make notes on paper, with timing going down one column, and with other columns including scene, some narrative, important dialogue and music. I'll draw a five-line staff and transcribe important themes. Often I will be rushing between the pause button, my notes, the piano, and my flute when making my timing tables.' [3]

Z obtains a lot of original material to be edited together later. In order to prepare his material, he relies on a more traditional methodology: he logs tape in his (digital) editing suite. He doesn't resent this often-laborious process (five to ten hours per hour of raw footage), as it gives him a chance to reflect upon the materials he has gathered. What he does welcome, however, is some way of automatically transcribing the speech from the video. As his subjects include non-native speakers and many interviews are conducted in the field (with field noise), his is a wish not likely to be realised in the near term.

Composite Scenario: Researchers who collect recordings for area studies, oral history and ethnographies often spend significant amount of time or money generating transcriptions, as illustrated by the following two examples:

'We used audio recording … the team included a stills photographer, we did not have funds for anything fancier and we felt comfortable and flexible with audio, because for example, we were often interviewing in public places where it would be difficult to use audiovisual recording such as cafes, pubs, hotels and noisy environments or in people's houses where people are five to a room … we used a cassette recorder … Since I am fluent in [the local language], virtually all interviews were conducted in [that language]; one or two subjects were more comfortable in another language and so one or two translators were sometimes used, making the interview a two or three-way conversation… In addition

to language issues, because not everyone is comfortable in [the researcher's language of choice], there are personal issues when interviewing traumatised people e.g. some women's voices descend to a whisper and become very difficult to hear. There is a very long period of time in terms of transcribing the tapes, because transcribers have to go backwards and forwards and backwards and forwards with people often speaking fast, or speaking in slang, and in mixed languages… Transcription is a very lengthy arduous process and I have to go over the transcripts, because some of the material gets garbled because transcribers put down words which are clearly not correct because they don't recognise them or haven't been able to make them out but I have handwritten interview notes that I can check things against and it is an extremely lengthy process to accuracy … It's a high-density activity when compared to going into an archive and taking notes on someone's letter, which you could either Xerox or translate if you are never going to see it again. It would be nice to have a package that I could feed in a tape and get a preliminary printout, but I feel this is very unlikely at present because of the [foreign] language issue…Once I have the transcript on the computer, I … can search for keywords and index everything according to key themes that I have drawn out … and these themes form the basis of an article.' [2]

'The resources I use are a mix of what historians would consider to be very traditional sources, archive sources such as looking at Colonial office papers, newspapers of the day, pamphlets, diaries, letters etc. The other part is using what is still considered non-traditional resources, which is oral history material. I will be going out and interviewing people who were involved in the period or living through [the period of interest]. I tape record everything. I choose audio recording for flexibility but it is also much less intimidating. The tape is unobtrusive. A typical interview lasts two or three hours and in some cases I will go back again and ask more questions, anything from two or three hours to six to eight hours of recorded material. The data can have an emotional range … you invariably stir up memories … the whole spectrum of emotions. And people get tired and children don't concentrate. And some interviews are recorded outdoors, so you've got wind etc. The other problem is corrugated iron roofs. Because of the roofs I can't interview when it rains, there are also tree frogs at night ... The tree frogs give you a hum all the way through, the air conditioning can be the same making nothing come out on the recording … Even if you're indoors the windows are open so you've got dogs barking or roosters crowing. I had one where this ruddy dog just didn't stop barking! Afterwards, and this is where I'm always looking for technology to help, there are two technical problems. Firstly transcribing, so a voice recognition system that could cope with [the local, highly accented] English would be brilliant but at the moment there is not anything sophisticated enough to do that as far as I know. And the second technical problem, trying to work with a qualitative database containing a very large number of interviews … but if you choose a sample right, the full set of interviews doesn't necessarily give you anything more than you'll get out of a smaller but more manageable subset of interviews. I use a professional transcriber – one of mine is very familiar now with [the local accent]. Then, in some sense, reading transcripts is no different from interpreting archive work.' [2]

Transcription problems are also faced by linguists, who develop detailed time-marked word or phonetic transcriptions for found data using tools such as wavesurfer or simply a pencil and paper:

'I'm studying recordings of English dialects … Transcription is difficult, word and phonetic with time alignment … locating the word boundaries is not an issue, because I can look at the spectrogram in wavesurfer, it is the transcription that takes time … When manual transcription is necessary, it is very expensive. Phonetic transcription speech takes 200-500 seconds of work to transcribe each second of speech. Manually transcribing a one-minute conversation, for example, can take one man-day. I would really like automatic transcription tools for word transcription and phonetic transcription at a fine phonetic level.' [3]

Film studies researchers working with foreign language data sometimes generate the same language or English language subtitles manually:

'Subtitling is very slow, done manually, so I would like software to transliterate the soundtrack into text – I'd be interested even if it gave a rough transcription. However, I think there'd be issues with [the foreign language's] dialects … There are probably 10 different dialects and the most commonly used words differ most widely across dialects with specialised words being more … trans-dialectal. And 80 to 90% of the data is colloquial at a guess.' [3]

Technology notes: Automatic transcription tools for specific situations do exist at commercial and research stages. These include systems for subtitling American and European television news, the English and Czech transcription tools developed by MALACH and phonetic transcription tools for conversation and American English, amongst many others. Such systems may provide first pass transcriptions that can be cleaned up in a shorter time than the time required for a full transcription but such assessments need to be made case-by-case. The sheer variety of languages and dialects, subject-specific vocabulary, recording environments, speech styles and emotions exhibited by humanities-relevant data (and hinted at by the descriptions above) means that an automated solution such as a general-purpose humanities relevant speech transcription server (e.g. a grid service) is rather unlikely to be feasible in the 2006-2010 timeframe although a potentially interesting, though lengthy, engineering research project might explore development of a transcription server designed for more constrained scenarios (e.g. a server for generating crude transcriptions of interviews in UK Southern English with participants wearing close-talking microphones). Such a server might share some similarities with the MIT lecture transcription server currently under development by one of their iCampus projects (MIT, 2006c).

Composite Scenario: X is a linguist interested in conversational analysis, Y studies performances and Z is a historian who might study speeches by a prominent political figure. Each of these can imagine research questions relating to the non-lexical content of a spoken word recording. They might benefit from certain types of automatic annotation, such as marking laughter, stress, pauses (and their durations) or emotional content. Such problems have been investigated by the engineering community but their solutions are apparently not readily available in a packaged form.

4.3.3 Analysis and interpretation

Scenario: X analyses films and television. He's learned that engineers have developed research tools which can automatically detect shot boundaries and can classify each shot into categories such as cuts or fades. 'With this technology I could explore questions such as the use of "long takes" or statistics about cuts and shot types and so on… I could extract statistics such as the number of cuts in the first and last 10 minutes of a film or historical changes in cutting rates in a TV or film type. It's too time-consuming to manually annotate these things for research on an extensive dataset … a tool giving this kind of quantitative analysis would be very useful.' [3] Such information may strengthen the empirical foundations of the kinds of research questions currently asked in the field, by providing quantitative evidence, but 'the most important part of the work will continue to be the interpretive analysis that explains why the statistics calculated should be so.' [3] He sees the envisaged tool as something facilitating existing research, but not engendering new types of research.

Composite Scenario: X works in theatre studies. He aims to take productions as a whole and to evaluate them in a much wider context, considering performance reception, sociocultural history, translation studies or details of the actual performances. He uses a variety of evidence about productions, including books, scripts, posters, theatre programmes, newspaper reviews etc. These resources come from theatre company archives and more general audiovisual archives such as the British Library Sound Archive. He also makes use of theatre-related broadband resources, but these are mostly for teaching rather than research. Videos or audio recordings of performances are also a possible type of evidence but access to performance recordings is not always straightforward. Although he is fortunate in having access to some audio recordings of performances and he's able to travel to archives which hold videotapes for some (though not all) productions in order to view them, more generally 'Colleagues in performing arts departments are crying out for a resource to allow visual records of performances to be widely accessible digitally for teaching and research purposes, but that is yet another kettle of fish with all the copyright implications […]' [4] He does note some interesting developments on the Web, such as the recent announcement of UK online pay-per-view theatre: 'This would be a major step forward.' [3]

The researcher sees possible uses for (manual or automatic) annotation tools for marking up interesting events in performances, but believes the likelihood of gaining access to annotate is quite unlikely for non-technological reasons. He says that rights and access issues place many limits on what he can achieve. For example, he believes that at present some researchers are benefiting from studying resources that are difficult for most people to access, because their work cannot easily be developed or critiqued. He feels this does not benefit the field. For example, he would like to perform comparative studies across different productions of the same play, across the different nights of a single production or even across rehearsals for that production and believes such analysis would significantly strengthen research in his area. Access to multiple productions would allow more detailed and quantitatively-supported investigation of casting decisions. For example, images from different productions can be used to compare the visual aspects of casting decisions at present, but easier access to performance videos

would allow comparisons of the different accents cast, such as when 'a particular character is played by a Scot or non-Scot.' [3]

The researcher has also learned that techniques from speech recognition can be used to time-align spoken words with text. These techniques might facilitate comparative studies if the right to work with the appropriate sources of video could be negotiated: a research project might explore the alignment of multiple production (audio or video) recordings to a master text, allowing easy comparison of how different parts of the text had been staged, lines modified, inordinately long deliveries, surprising or strange uses of words and the uses of pauses.

This researcher is also thinking about other ways of gathering research information from different productions, such as a multimedia wiki: he envisages a web site that allows productions to submit their own recordings, e.g. school productions, and maintains access controlled sections to allow commentary and annotations from academics, schools or others.

Scenario: X is a theologian. He comments that his is a highly interdisciplinary subject 'in the sense that for any other given [arts and humanities] discipline, there will be a branch of theology and religion that is similar. Thus, the humanities disciplines that have uses for the tools we are looking at will give hints of the potential uses in theology. For example, if the practice based arts have a use for them, I can find an equivalent in theology. Consider performance: then theologians will consider performance within the context of ritual.' [2] Thus, the tools suggested in the previous scenario (and elsewhere) may find more general application than may be immediately apparent.

4.3.4 Dissemination

Composite Scenario: X is an area studies researcher and his current project is compiling a collection of clips containing 'interesting' objects and events in the moving images and/or the soundtrack. He will deposit his annotations and perhaps the clips in a digital repository at the end of the project (such as a research council or university repository). He believes the repository's current text catalogue will not encourage reuse of his dataset, because textual descriptions do not capture the richness of the timebased audiovisual content; in contrast, an interface which supported audiovisual browsing or even searching of the spoken word content would be more appropriate for encouraging the use of this kind of data.

Many researchers collect interviews e.g. theatre researchers interviewing set designers, oral historians interviewing the target group and ethnoarchaeologists interviewing local people close to a dig or field survey site. They also suggest that new technologies could be used to make their data collections more accessible and thereby encourage reuse. For example, the spoken word content could be indexed based upon (automatically time-aligned) transcripts, if available, or based upon less accurate but automatically-generated and time aligned transcripts. This was perceived as a possible improvement over a very limited catalogue description of their collections. Such indexing systems are feasible for languages for which speech-to-text functionality and pronunciation dictionaries exist, though performance

would vary depending upon the speech-to-text system quality and the careful design of the associated time-alignment algorithm.

Scenario: X publishes material that involves commercial, popular culture. She is in the process of writing a book chapter about a recent, fairly well-known art-film. She wants to include a still from the film as an illustration in the book, but had to dedicate months of requests, follow-ups, and negotiations to include the one still, eventually at the price of £700. She doesn't resent that, as she often budgets for such contingencies. What is really frustrating, she finds, are the intellectual property owners who don't even acknowledge such requests.

4.3.5 Other uses

Scenario: X is a linguist and is able to spend a reasonable amount of time either self-recording research data or in negotiating data release for research purposes. However, this is not a luxury afforded to students working on class projects or short pieces of research. The ability to search for spoken word audio or video online may be a mechanism by which students can rapidly construct small data sets for research purposes e.g. a selection of UK broadcast media coverage about the Siege of Beslan. 'It's not always a solution, because some research questions will require more demographic information about the participants in the audio than is available from the source websites, but this can also be a problem with data obtained from traditional audiovisual archives … For some research questions it could significantly speed up the data collection process and free up time for additional research.' [3]

Scenario: X often has to respond to questions about language usage from the media and the general public. Spoken word examples are sometimes more appropriate than examples based on written text. Audio and video search engines may provide a convenient mechanism for locating illustrative clips, whether in contemporary produced video or in the self-recordings individuals have uploaded to the Web.

Scenario: X is a linguist who occasionally uses clips from a standard linguistics database to illustrate certain phonological changes exhibited by specific languages or dialects for students and lecturers. The ability to search for spoken word audio on the Web may provide an alternative source of examples. However, getting the equipment together to play audiovisual clips in some of the older lecture halls can be so time-consuming that it acts as a deterrent to actually playing these clips to class, though he might give the appropriate links to the student in a handout.

Composite Scenario: X works in a modern languages department. He observes that language teachers regularly make off-air recordings on third-generation videotapes and then spend considerable time watching them to find relevant sections for class or for use in comprehension tests. A package which could make these recordings searchable would be very useful. It would ideally support many foreign languages. Even a mechanism which filtered out non-spoken word audio might speed up the process of browsing the spoken content for relevant phrases, topics or linguistic contexts.

Scenario: X teaches a course on the late 20th century history of an English-speaking country. 'A technology like Google for video would be very useful to me, for teaching as well as anything else. How far

back does the data currently online go, as far as the 1960s-1980s? I might also look for relatively recent reports on political activity in the area.' [2]

Scenario: X and Y are archaeologists who report that the use of video is very common for making trench recordings. (These show information such as soil changes or the relationship between points in the trench better than individual images.) They believe a significant amount of time is spent editing such recordings down to a final annotated set, in the evening or some other later date. They are curious about the possibilities for using speech technology (or some other technology) for supporting annotation of images and video in the field, envisaging a system linked to the large cameras used in excavations or the smaller portable cameras used by surveyors. 'Perhaps this would obviate the need to edit down the video collected: could one simply search the soundtrack for the relevant spoken annotation in order to locate the segment of interest, such as "trench layer 6"?' [3] This is an interesting problem for exploration from a speech recognition (or, more generally, the mobile computing) point-of-view because the situation can be controlled in several ways e.g. these two researchers 'would be willing to wear close talking microphones to reduce noise,' [3] there are limits upon the number of speakers (i.e. the set of researchers who would be digging) and it seems likely there exist constraints upon the annotation vocabulary and grammar that could be usefully exploited by technology solutions.

4.4 Technical expectations

Beyond the questions about general audiovisual usage, we observed a few common misapprehensions in the interviews. They involved mismatched expectations about the likely performance of 'black box' tools for AV in envisaged deployments.

4.4.1 Error

Some interviewees tended to assume that automatic annotation systems would provide correct and unarguable results. For example, one suggestion that arose several times was to use an automatic speech-to-text system to generate a *perfect* AV concordance tool that would allow a researcher to examine *all* of the AV contexts in which specific words appear. This suggestion could be challenging to solve today because most automatic annotation techniques operate with some degree of error and thus a concordance based on a speech-to-text transcript is not guaranteed to be perfect. Similarly, an automatic phonetic transcription system or shot boundary detection system will not always yield the same results as a careful and experienced human, and a system for searching video based on an automatic speech-to-text transcription of the soundtrack is not guaranteed to yield the same results as a system that searches a human transcription of the soundtrack. In some cases the apparent disagreement between automated system and human results may be due to genuine ambiguity in the data, which could be revealed by a comparison of the disagreement between multiple humans performing the same task; in other cases, errors arise due to limitations of the automatic system such as a lack of robustness (see next). The level of

error which could be tolerated in an automatic annotation system will generally be application- and data-set specific.

4.4.2 Robustness

Many of the techniques discussed in this report make use of statistical methods. These methods extract information from a set of 'learning' or 'training' examples and make use of this information when analysing new examples. Since this set of training examples is usually incomplete, the information extraction process leads to a system which is often (though not always) 'brittle' when challenged by new examples that are different from those it has seen before (i.e. it is not *robust* in the face of new examples). This means statistically-based techniques cannot always be used as 'black boxes' that can be expected to operate equally well (i.e. at equal error rate levels) on all types of new examples: most deployments must investigate tuning or 'adaptation' techniques to make the base systems perform better in the new deployment scenario.

For example, a speech recognition system developed using examples of carefully dictated, Received Pronunciation speech may perform quite acceptably when applied to carefully dictated, Received Pronunciation speech. However, in many if not most situations, the same system may struggle when faced with more diverse types of speech such as conversational speech or highly accented slang, sometimes even after adaptation. Even apparently similar speech can sometimes be problematic, as illustrated by the lack of robustness of a system trained on broadcast new shows when applied to speech from a news show the system hasn't seen before. At present, the ideal is to use a base system trained on data representative of the data which will be seen during deployment and to perform additional system adaptation.

4.4.3 A lack of appreciation for the *demo effect*

It is well accepted within technical communities that a new system will be shown in the best possible light in terms of input and performance. Translating such a demonstration to an applied research situation can cause a shocking disappointment to researchers unprepared for it. It is for this reason that we label the maturity of the technologies mentioned in appendix B.

4.4.4 'I can't do that [with that tool]'

A common complaint was also related to the usability of increasingly complex software and devices. One interviewer repeatedly and enthusiastically pointed out features and capabilities of media applications that he and the researchers mutually used. One music researcher shied away from using iTunes or her iPod when giving talks because she didn't believe she could easily set up extracts in iTunes or get random access within a track on an iPod. Although they are not common functions, both are possible with a little investigation. It is often simply a question of knowing that something is possible with common software.

5 References

Adams, B., Lin, C.-Y. & Iyengar, G. (2002). IBM Multimodal Annotation Tool. http://www.alphaworks.ibm.com/tech/multimodalannotation

AHDS, Arts and Humanities Data Service (2006a). Creating Digital Audio Resources. http://www.ahds.ac.uk/creating/guides/audio-resources/GGP_Audio_Contents.htm

AHDS, Arts and Humanities Data Service (2006b). Enabling Digital Resources for the Arts and Humanities. http://ahds.ac.uk/

AHDS Performing Arts (2005). Designing Shakespeare – an AHDS Performing Arts collection. http://ahds.ac.uk/performingarts/collections/designing-shakespeare.htm

AHRC, Arts and Humanitites Research Council (2003). Research subject coverage. http://www.ahrc.ac.uk/about/subject_coverage/research_subject_coverage.asp

AHRB Centre for British Film and Television Studies (2005). Moving History. http://www.movinghistory.ac.uk/archives/se/films/se5margate.html

Allan, J. (2001). Perspectives on information retrieval and speech. Lecture Notes In Computer Science; Vol. 2273, 1–10. http://ciir.cs.umass.edu/pubfiles/ir-236.pdf

Allan, J. (2003). Robust techniques for organising and retrieving spoken documents. *EURASIP Journal on Applied Signal Processing*, 103–114. DOI: 10.1155/S1110865703211070. http://www.hindawi.com/GetArticle.aspx?doi=10.1155/S1110865703211070

Altavista (2006). http://www.altavista.com/

Amatriain, X., Massaguer, J., Garcia, D. & Mosquera, I. (2005). The CLAM Annotator: A Cross-platform Audio Descriptors Editing Tool. *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 426–429. http://www.iua.upf.edu/mtg/publications/9317d2-ismir2005-clam-annotator.pdf

Amir, A., Srinivasan, S. & Efrat, A. (2002). Search the Audio, Browse the Video — A Generic Paradigm for Video Collections. http://www.hindawi.com/GetArticle.aspx?Doi=10.1155/S111086570321012X&e=CTA

Annodex (2006). http://www.annodex.net/

Archaeological computing (1996). Computer science and the management of an archaeological excavation. *Archaeological Computing Newsletter, 50*, 6 May 1996, 13.

Arons, B (1997). SpeechSkimmer: A System for Interactively Skimming Recorded Speech. *ACM Transactions on Computer-Human Interaction*, *4*, 3–38, http://xenia.media.mit.edu/~barons/html/tochi97.html

Atwood, M. (2006). The Immigrants. http://www.poetryarchive.org/poetryarchive/singlePoem.do?poemId=97

Aurix (2006). Phonetic audio mining, speech recognition software. http://www.aurix.com/product/index.htm

Autonomy (2006). Understanding the hidden 80%. http://www.autonomy.com/content/home/

BBC, British Broadcasting Corporation (2003a). Creative Licence Group. http://creativearchive.bbc.co.uk/

BBC, British Broadcasting Corporation (2003b). The Full Licence. http://creativearchive.bbc.co.uk/licence/nc_sa_by_ne/uk/prov/

BBC, British Broadcasting Corporation (2006). BBC Programme Catalogue. http://open.bbc.co.uk/catalogue/infax

BBN Technologies (2004–06a). Audio Indexing System. http://www.bbn.com/Products_and_Services/Unstructured_Data/Audio_Indexing_System.html

BBN Technologies (2004–06b). IdentiFinder. http://www.bbn.com/Products_and_Services/Unstructured_Data/Identifinder.html

BEPC, British Electronic Poetry Centre (2004). http://www.soton.ac.uk/~bepc/

Besser, H. (2001). Digital Preservation of Moving Image Material? *The Moving Image,* Fall 2001. http://www.gseis.ucla.edu/~howard/Papers/amia-longevity.html

BFI, British Film Institute (2002). British Empire in colour. http://www.bfi.org.uk/features/interviews/empireincolour.html

Bignell, J. (2005). Exemplarity, Pedagogy and Television History. *New Review of Film and Television Studies*, *3(1)*, 15–32.

Blinkx (2006). blinkx.tv http://tv.blinkx.com/

Bosma, M., R.C. Veltkamp & F. Wiering (2006). Muugle: a music retrieval experimentation framework. In M. Baroni, A. R. Addessi, R. Caterina, M. Costa (2006) *Proceedings of the 9th International Conference on Music Perception & Cognition*, 1297–1303. http://www.give.nl/multimedia/publications/pdf/icmpc06.pdf

British Academy (2005). E-Resources for Research in the Humanities and Social Sciences – A British Academy Policy Review. http://www.britac.ac.uk/reports/eresources/index.html

British Library (2006a). Intellectual property: a balance; the British Library manifesto, http://www.bl.uk/news/pdf/ipmanifesto.pdf

British Library (2006b). The British Library Sound Archive. http://www.bl.uk/collections/sound-archive/cat.html

British Library (2006c). Collect Britain. http://www.collectbritain.co.uk/collections/dialects/

British Library (2006d). Oral history: Millennium Memory Bank. http://www.bl.uk/collections/sound-archive/millenni.html

British Library (2006e). Classical Music. http://www.bl.uk/collections/sound-archive/wam.html

BUFV, British Universities Film and Video Council (2002). Hidden Treasures Conference. http://www.bufvc.ac.uk/publications/articles/Treasures.pdf

BUFVC, British Universities Film and Video Council (2004). Hidden Treasures: the UK Audiovisual Archive Strategic Framework. http://www.bufvc.ac.uk/faf/HiddenTreasures.pdf

BUFVC, British Universities Film and Video Council (2005). Researcher's Guide Online. http://joseph.bufvc.ac.uk/RGO/index.html

BUFVC, British Universities Film and Video Council (2006). Moving Image Gateway. http://www.bufvc.ac.uk/gateway/

Byrne, B. (2006). Personal communication to Harriet Nock.

Carnegie Mellon University (1994–2006). Informedia Digital Video Library, http://www.informedia.cs.cmu.edu/

Carnegie Mellon University (2005a). The ESP Game: Labelling the Web. http://www.espgame.org/

Carnegie Mellon University (2005b). Peekaboom. http://www.peekaboom.org/

Carson, C. (2005). Digitising Performance History: Where Do We Go from Here? *Performance Research*, *10(3)*, 4–17.

Cee, D., Emele, M. & Spree, L. (1996). Video-recording as part of the critical archaeological process. http://www.catalhoyuk.com/TAG_papers/karlsruhe1.htm

Centre for Digital Music, Queen Mary, University of London (n.d.). Sonic Visualiser. http://www.sonicvisualiser.org/

Centre for Spoken Language Understanding, OGI School of Science & Engineering, Oregon Health & Science University (n.d.). CSLU Toolkit. http://cslu.cse.ogi.edu/toolkit/

Chalmers, A., Stoddart, S., Belcher, M. & Day, M. (1996). An Interactive Photo-Realistic Visualisation System for Archaeological Sites. http://www.cs.bris.ac.uk/~alan/Arch/INSITE/research/comvis/insite2.htm

Chan, H.Y. & Woodland, P. (2004). Improving broadcast news transcription by lightly supervised discriminative training. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004*, *1*, 17–21. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=9248, Digital Object Identifier 10.1109/ICASSP.2004.1326091. Preprint: http://mi.eng.cam.ac.uk/reports/svr-ftp/chan_icassp2004.pdf

Chen, S. (2006). Personal communication to Harriet Nock.

Church, K. W. (2003). Speech and Language Processing: Where Have We Been and Where Are We Going? *Proceedings of EUROSPEECH 2003, 8th European Conference on Speech Communication and Technology*, Geneva, 1–4. http://research.microsoft.com/users/church/wwwfiles/papers/Eurospeech/2003/ES032000.pdf

CNN, Cable News Network (2006). Image Source. http://www.cnnimagesource.com/CNIS/index.html

Connolly, C. (2004). Copyright issues relating to Film Studies and Modern Languages staff and students. Seminar *The tangled web – making sense of copyright in developing and exploiting on-line resources*, CILT, London, 24 November 2004. http://www.lang.ltsn.ac.uk/resourcedownloads.aspx?resourceid=2060&filename=connolly.rtf

Cowie, J. & Wilks, Y. (2000). Information extraction. In R. Dale, H. Moisl and H. Somers (eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker. http://www.dcs.shef.ac.uk/~yorick/papers/infoext.pdf

Creative Commons (2006). Enabling the legal sharing and reuse of cultural, educational, and scientific works. http://creativecommons.org/

Cyberlink (2004). CyberLink PowerDVD Ups Protection. http://www.prnewswire.co.uk/cgi/news/release?id=114946

Czerwinski, M. Gage, D. & Gemmell, J. (2006). Digital memories in an era of ubiquitous computing and abundant storage. *Communications of the ACM, 49(1)*.

Devon Technologies (2006). DEVONthink. http://www.devon-technologies.com/products/devonthink/index.html

DiMeMa (2006). CONTENTdm. http://www.dimema.com/products/overview.html

Dixon, S. & Widmer, G. (2005). MATCH: a music alignment tool chest. *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*, London, 492–497. http://www.ofai.at/cgi-bin/tr-online?number+2005-17

Dixon, S. (2005). MATCH, Music Alignment Tool Chest. http://www.ofai.at/~simon.dixon/match/index.html

Dunn, J. W., Byrd, D., Notess, M., Riley, J. & Scherle, R. (2006). Variations2: Retrieving and using music in an academic setting. *Communications of the ACM, 49(8)*, 52–58. http://delivery.acm.org/10.1145/1150000/1145314/p53-dunn.pdf?key1=1145314&key2=4840899511&coll=&dl=acm&CFID=15151515&CFTOKEN=6184618

Edina (2006). Film and Sound Online. http://www.filmandsound.ac.uk

eGovMonitor (2005). Interview – Newsfilm online, digitising ITN and Reuters archives. http://www.egovmonitor.com/node/2240

Fasttalk (2006). http://fasttalk.com/

Ferne, T. (2005) Annotatable audio revisited. http://cookinrelaxin.blogspot.com/2005/12/annotatable-audio-revisited.html

Fiscus, J. G., Radde, N., Garofolo, J., Le, A., Ajot, J. & Laprun, C. (2005). The Rich Transcription 2005 Spring Meeting Recognition Evaluation. http://www.nist.gov/speech/publications/papersrc/rt05sresults.pdf

Fitzgerald, R. A. (2003). Performer- dependent dimensions of timbre: identifying acoustic cues for oboe tone discrimination. PhD Thesis, School of Music, University of Leeds.

Fox, J.C. (2003). Dr. J.C. Fox: Filming women in the Third Reich. http://www.dur.ac.uk/~dhi0www/staff/jcf.html

Franz, M., Ramabhadran, B., Ward, T. & Picheny, M. (2003). Automated Transcription and Topic Segmentation of Large Spoken Archives. *Proceedings of Eurospeech 2003*. http://www.clsp.jhu.edu/research/malach/pubs/euro03-ir.pdf

FreeSound (2006). The Freesound Project. http://freesound.iua.upf.edu/

Furui, S. (2005a). 50 years of progress in speech and speaker recognition. *Proceedings of SPECOM 2005*, Patras, Greece, 1–9. Preprint: http://www.furui.cs.titech.ac.jp/publication/2005/SPCOM05.pdf

Furui, S. (2005b). Spontaneous speech recognition and summarization, The Second Baltic Conference on Human Language Technologies, 39–50. http://www.furui.cs.titech.ac.jp/publication/2005/HLT2005.pdf

Gales, M. J. F. (1996). The generation and use of regression class trees for MLLR adaptation. *Technical Report CUED/F-INFENG/TR263*, Cambridge University, 1996. http://citeseer.ist.psu.edu/gales96generation.html

Games-db (2006). http://www.games-db.com/

Gauch, S., Li, W. & Gauch, J. (1997). The VISION Digital Video Library. http://www.ittc.ku.edu/publications/documents/Gauch1997_IPM97.pdf

Gauvain, J.-L., & Lamel, L. (2003). Structuring Broadcast Audio for Information Access. *EURASIP journal on Applied Signal Processing*, 2003 (2): 140–150. http://hindawi.com/GetArticle.aspx?doi=10.1155/S1110865703211033

Goldman, J., Renals, S., Bird, S., de Jong, F., Federico, M., Fleischhauer, C. et. al. (2005). Accessing the spoken word. *International Journal on Digital Libraries*, *5(4)*, 287–298. DOI: 10.1007/s00799-004-0101-0. Preprint: http://www.cstr.ed.ac.uk/downloads/publications/2005/swag-ijdl05.pdf, full project report: http://www.dcs.shef.ac.uk/spandh/projects/swag/swagReport.pdf

Gomez, E. & Bonada, J. (2005). Tonality Visualization of Polyphonic Audio, *Proceedings of the International Computer Music Conference*, Barcelona, 57–60.

Gonet, W. & Święciński, R. (2002). Speech Lab @ Work and @ Home.*Speech and Language Technology*, *6*, 57–80, Polish Phonetic Association.

Google (2005). Searching for Music. http://googleblog.blogspot.com/2005/12/searching-for-music.html

Google (2006a). Nara on Google Video. http://video.google.com/nara.html

Google (2006b). Google Video. http://video.google.co.uk/

Google (2006c). Google Video Upload program. https://upload.video.google.com/

Google (2006d). Video Upload Program for major producers. https://services.google.com/inquiry/video

Goto, M. & Goto, T. (2005). Musicream: New Music Playback Interface for Streaming, Sticking, Sorting and Recalling Musical Pieces. *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 404–411. http://ismir2005.ismir.net/proceedings/1058.pdf

Grabe, E., Nolan, F. & Post, B. (2003). English intonation in the British Isles. http://www.phon.ox.ac.uk/IViE/

Gracenote (2006). Gracenote Music Fans. http://www.gracenote.com

Grishman, R. (1997). Information extraction: techniques and challenges. In Maria Teresa Pazienza (ed.) *Information Extraction*. Springer-Verlag, Lecture Notes in Artificial Intelligence. http://citeseer.ist.psu.edu/grishman97information.html; http://nlp.cs.nyu.edu/muc/ie-survey-frascati-97.ps

Grishman, R. (1998). Information extraction and speech recognition. *Proceedings of the DARPA News Transcription and Understanding Workshop*. http://citeseer.ist.psu.edu/ralph98information.html; http://nlp.cs.nyu.edu/publication/papers/bntuw98.ps

Gustman S., D. Soergel, D. Oard, W. Byrne, M. Picheny, B. Ramabhadran & D. Greenberg (2002). Supporting access to large digital oral history archives. *Joint Conference on Digital Libraries*, Portland, http://www.clsp.jhu.edu/research/malach/pubs/JCDL2002MALACH.pdf).

Hansen, J.H.L., J.R. Deller, Jr., & M.S. Seadle (2001). Engineering Challenges in the Creation of a National Gallery of the Spoken Word: Transcript-Free Search of Audio Archives. *Joint Conference on Digital Libraries* (IEEE & ACM – JCDL 2001), Roanoke, VA, http://cslr.colorado.edu/beginweb/ngsw/publications/CP-JCDL-NGSW.Jun01.pdf

Hauptmann, A. (2005). Lessons for the Future from a Decade of Informedia Video Analysis Research. http://www.informedia.cs.cmu.edu/documents/CIVR05_Hauptmann.pdf

Herrera, P., Celma, O., Massaguer, J., Cano, P., Gomez, E., Gouyon, F. et. al. (2005). MUCOSA: A Music Content Semantic Annotator. *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 77–83. http://ismir2005.ismir.net/proceedings/1115.pdf

Howard-Spink, S. (n.d.). You just don't understand!. http://domino.watson.ibm.com/comm/wwwr_thinkresearch.nsf/pages/20020918_speech.html

HTK (n.d.). HTK Speech Recognition Toolkit. http://htk.eng.cam.ac.uk/

Humaine (2003–06). Related Projects. http://emotion-research.net/wiki/RelatedProjects

HUMBUL (2006). Intute: Arts and Humanities. http://www.intute.ac.uk/artsandhumanities/langlit-all/

Huron, D. (2002). The Humdrum Toolkit. http://csml.som.ohio-state.edu/Humdrum/

IBM, International Business Machines (2006). Marvel. http://domino.research.ibm.com/comm/research_projects.nsf/pages/marvel.index.html

IBM, International Business Machines (n.d.). Speech-to-Speech Translation. http://domino.watson.ibm.com/comm/research.nsf/pages/r.uit.innovation.html

IDIAP Research Institute (2006). Ferret meeting browser demo. http://mmm.idiap.ch/demo/

IMDb (2006). The Internet Movie Database. http://www.imdb.com/

Imperial War Museum, (2006a). IWM Collections Online. http://www.iwmcollections.org.uk/

Imperial War Musuem, (2006b). War on Land. http://www.iwmcollections.org.uk/onland/essay.asp

Indiana University (2005). Variations2: The Indiana University Digital Music Library. http://www.dml.indiana.edu/

Information Systems Research Laboratory, University of Illinois (2005), M2K (Music-to-Knowledge): A tool set for MIR/MDL development and evaluation. http://www.music-ir.org/evaluation/m2k/

Institute for Multimedia Literacy, University of Southern California (2006). MediaBASE. http://www.iml.annenberg.edu/html/research/mediabase.htm

InternetArchive (2006). About the Movies. http://www.archive.org/about/faqs.php#About_the_Movies

Isaacson, E. (2005). What You See is What You Get: On Visualizing Music. *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 389–395. http://ismir2005.ismir.net/proceedings/1129.pdf

ISMIR (n.d.). The International Conferences on Music Information Retrieval and Related Activities. http://www.ismir.net/

JISC (2005a). Press Release: JISC, BUFVC and ITN to work in partnership. http://www.jisc.ac.uk/index.cfm?name=pr_itn_bufvc_itn_partnership

JISC (2005b). Digital Repositories: Helping universities and colleges. http://www.jisc.ac.uk/uploaded_documents/HE_repositories_briefing_paper_2005.pdf

JISC (2006). JISC Digitisation Program. http://www.jisc.ac.uk/digitisation_home.html

Kelly, C., Byford, A. & Jones, P. (2004). Childhood in Russia 1890–1991: a Social and Cultural History. http://www.mod-langs.ox.ac.uk/russian/childhood/

Kim, D.Y., Chan, H.Y., Evermann, G., Gales, M., Mrva, D. Sim, K. & Woodland, P.C. (2005). *Development of the CU-HTK 2004 Broadcast News Transcription Systems. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005 (ICASSP '05), *1*, 861–864. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=9711, Digital Object Identifier 10.1109/ICASSP.2005.1415250. Preprint: http://mi.eng.cam.ac.uk/reports/svr-ftp/kim_icassp05.pdf

Kinder, M. (n.d). The Labyrinth Project. http://www.annenberg.edu/labyrinth/

Klapuri, A. (2004). Automatic music transcription as we know it today. *Journal of New Music Research*, *33*, 269–282.

Koumpis, K. & Renals, S. (2001). The role of prosody in voicemail summarization systems. *ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ. http://www.dcs.shef.ac.uk/~sjr/pubs/2001/pros01-vm.html

Koumpis, K. & Renals, S. (2005). Content-based access to spoken audio. *IEEE Signal Processing Magazine*, *22(5)*, 61–69. Preprint: http://www.cstr.ed.ac.uk/downloads/publications/2005/koumpis-spm05.pdf

Kraaij, W., Smeaton, A., Over, P. & Arlandis, J. (2004). TRECVID 2004 – An Overview. http://www.cdvp.dcu.ie/Papers/TRECVID2004_Overview.pdf

KTH (n.d.). The NICO Toolkit. http://nico.sourceforge.net/

Le, A. (2004). 2004 Fall Rich Transcription Speech-to-Text Evaluation. http://www.nist.gov/speech/tests/rt/rt2004/fall/rt04f-stt-results-v6b.pdf

Leath, T. (2005). Documents. http://www.infm.ulst.ac.uk/~ted/html/papers.htm

Lee, L. & Chen, B. (2005). Spoken document understanding and organization. *IEEE Signal Processing Magazine*, *22(5)*, 42–60. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?isnumber=32367&arnumber=1511823&count=17&index=4

Lesaffre, M., Leman, M., De Baets, B. & Martens, J.-P. (2004). Methodological considerations concerning manual annotation of musical audio in function of algorithm development. *Proceedings of the Fifth International conference on Music Information Retrieval*, Barcelona, pp. 64–71. http://www.ipem.ugent.be/MAMI/Public/Papers/ISMIR2004_LesaffreEtALFinal.pdf

Linguistic Data Consortium (1996–2005). Global Autonomous Language Exploitation. http://projects.ldc.upenn.edu/gale/

Linguistic Data Consortium (2001). Linguistic Annotation. http://www.ldc.upenn.edu/annotation/

Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Ang, J., Hillard, D., et. al. (2005). Structural metadata research in the EARS program. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings* (ICASSP '05), *5*, 18–23. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=9711, Digital Object Identifier 10.1109/ICASSP.2005.1416464. Preprint: http://www.icsi.berkeley.edu/~yangl/icassp2005-mde.pdf

Llisterri, J. (2006). Speech analysis and transcription software. http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html

Logan, B., Moreno, P. & Van Thong, J.M. (2003). Approaches to Reduce the Effects of OOV Queries on Indexed Spoken Audio. *Technical report HPL-2003-46*, HP Laboratories Cambridge. http://citeseer.ist.psu.edu/logan03approaches.html

Longuet-Higgins, H.C. (1976). Perception of melodies. *Nature*, *263*, 646–653.

Luscombe, S. (2002). The British Empire in colour. http://www.britishempire.co.uk/media/documentary/britishempireincolour.htm

Lyman, P. & Varian, H. (2003). How Much Information? http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Machine Intelligence Laboratory, University of Cambridge (2005). AGILE: Autonomous Global Integrated Language Exploitation. http://mi.eng.cam.ac.uk/research/projects/AGILE/

MALACH (n.d.). MALACH: Multilingual Access to Large Spoken Archieve. http://www.clsp.jhu.edu/research/malach/

Marchionini, G. & Geisler, G. (2002). The Open Video Digital Library. *D-Lib Magazine*, *8(12)*. http://www.dlib.org/dlib/december02/marchionini/12marchionini.html

Martin, A., & Przybocki, M. (2003). NIST 2003 language recognition evaluation. *Proceedings of EUROSPEECH 2003, 8th European Conference on Speech Communication and Technology*, Geneva, 1341–1344. http://www.nist.gov/speech/publications/papersrc/eurospeech_03_final.pdf

MeeVee (2006). http://www.meevee.com/

Meredith, D. (2006). The *ps13* Pitch Spelling Algorithm. *Journal of New Music Research*, *35*, 121–159.

Metavid (2006). About Metavid. http://metavid.ucsc.edu/

MIC (2006). Moving Image Collection. http://mic.imtc.gatech.edu/public_portal/public_collectionsexplore.php

Michigan State University (2005). MediaMatrix: Isolate, segment and annotate video. http//www.matrix.msu.edu/~mmatrix/

Microsoft (2006). Speech SDK 5.1. http://www.microsoft.com/downloads/details.aspx?FamilyId=5E86EC97-40A7-453F-B0EE-6583171B4530&displaylang=en

Middleton, P. (2003). How to Read a Poetry Reading: Reading the Reading. http://www.soton.ac.uk/~bepc/forum/middleton_readingessay.htm

Mills, S. (2004). Auditory archaeology at Çatalhöyük: preliminary research. http://www.catalhoyuk.com/archive_reports/2004/ar04_40.html

MIREX (n.d.). http://www.music-ir.org/mirexwiki/index.php/Main_Page

MIT, Massachusets Institute of Technology (2006a). Welcome to DSpace. http://www.dspace.org

MIT, Massachusets Institute of Technology (2006b). OpenCourseware Initiative. http://icampus.mit.edu/projects/DSpace.shtml

MIT, Massachusets Institute of Technology (2006c). Project: Spoken Lecture Processing. http://icampus.mit.edu/projects/SpokenLecture.shtml

Miyamori, H., Stejic, Z., Araki, T., Minakuchi, M. & Ma, Q. (2006). Proposal of Integrated Search Engine of Web and TV Contents. *Proceedings of WWW2006, 15th World Wide Web Conference*, Edinburgh. http://www2006.org/programme/files/pdf/p190.pdf

Moore, R., (2003). A Comparison of the data requirements of automatic speech recognition systems and human listeners. *Proceedings of EUROSPEECH 2003, 8th European Conference on Speech Communication and Technology*, Geneva, 2582–2584. http://www.dcs.shef.ac.uk/~roger/publications/Eurospeech03%20Comparison%20of%20Data%20Requirements.pdf

Mörchen, F., Ultsch, A., Nöcker, M. & Stamm, C. (2005). Databionic Visualization of Music Collections According to Perceptual Distance, *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 396–403. http://www.mybytes.de/papers/moerchen05ismir.pdf

MTG, Pompeu Fabra University (2006). http://iua-share.upf.es/wikis/clam/index.php/Music_Annotator

National Gallery of the Spoken Word (n.d.). http://www.ngsw.org/

Naxos (2006). Naxos music library. http://www.naxosmusiclibrary.com

NCeSS, National Centre for e-Social Science (2005). MixedMediaGrid. http://www.ncess.ac.uk/nodes/mimeg/

NCHSwiftSound, (2006). Express Scribe Transcription Playback Software. http://www.nch.com.au/scribe/index.html

net imperative (2006). Blinkx launches ad-funded video service. http://www.netimperative.com/2006/02/08/Blinkx_ITN

News.com (2006). Google puts National Archives video online. http://news.com.com/Google+puts+National+Archives+video+online/2100-1025_3-6043193.html

Nixon, L. & Price, S. (2000). Sphakia Survey: The internet edition. http://sphakia.classics.ox.ac.uk/video.html

Nixon, L. & Price, S. (2004). Paper, Video, Internet: New Technologies for Research and Teaching in Archaeology: The Sphakia Survey. *Journal of Interactive Media in Education* (Designing and Developing for the Disciplines Special Issue), 2004 (17). ISSN:1365-893X. http://www-jime.open.ac.uk/2004/17

Northwestern University (2006). Project Pad. http://projectpad.northwestern.edu/ppad2/

NYU, New York University (n.d.). Query by Humming. http://querybyhum.cs.nyu.edu/

Oard, D., Demmer-Fushman, D. & Hajic, J. (2002). Cross-Language Access to Recorded Speech in the MALACH Project. ftp://ftp.umiacs.umd.edu/pub/bonnie/tsd2002.pdf

Oard, D. et al. (2004). Building an information retrieval test collection for spontaneous conversational speech. *27th Annual International ACM SIGIR Conference (SIGIR2004)*, Sheffield. http://papers.ldc.upenn.edu/SIGIR2004/IR.pdf

Open Directory Project (2006). Games:Video Games: History. http://dmoz.org/Games/Video_Games/History/

OpenArchives (2006). Open Archives Initiative. http://www.openarchives.org/

OpenPress (2005). UKTN to Stage World's First Trial of Pay-Per-View Theatre. http://www.theopenpress.com/index.php?a=press&id=5160

OpenVideo (2005). The Open Video Project. http://www.open-video.org/

Ostendorf, M., Shriberg, E., & Stolcke, A. (2005). Human language technology: opportunities and challenges. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005. (ICASSP '05), *5*, 18–23. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=9711, Digital Object Identifier 10.1109/ICASSP.2005.1416462. Preprint: http://www.speech.sri.com/papers/icassp2005-specialsession.pdf

Pardo, B. & Birmingham, W.P. (2003). Query by humming: how good can it get? In J.S. Downie (ed.) *The MIR/MDL Evaluation Project White Paper Collection*, Edition #3, 107–109. http://www.music-ir.org/evaluation/wp3/wp3_pardo_query.pdf

PC Magazine (2006). IBM Strives for "Superhuman" Speech Tech, 24 January 2006. http://www.pcmag.com/article2/0,1895,1915071,00.asp

Peeters, G., La Burthe, A. & Rodet, X. (2002). Toward Automatic Music Audio Summary Generation from Signal Analysis. *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Paris. http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2002_ISMIR_AudioSummary.pdf

Pickering, M. (2006). Automatic news summarization extraction system. http://wwwhomes.doc.ic.ac.uk/~mjp3/anses/

Pickering, M.J., Wong, L. & Rüger, S.M. (2003). ANSES – Summarisation of news video. *Proceedings of International Conference on Image and Video Retrieval (CIVR-2003)*. Lecture Notes in Computer Science 2728 (Springer Verlag), 425–434. http://www.doc.ic.ac.uk/~mjp3/phd/www-pub/civr2003.pdf

Plichta, B. & Kornbluh, M. (n.d.). Digitizing speech recordings for archival purposes. http://www.historicalvoices.org/papers/audio_digitization.pdf

Potamianos, G., Neti, C., Gravier, G. Garg, A., & Senior, A.W. (2003). Recent Advances in the Automatic Recognition of Audio-Visual Speech, *Proceedings of the IEEE*, *91*. http://www.research.ibm.com/AVSTG/IEEEPROC03_REVIEW.pdf

Praktische Informatik IV, University of Mannheim (2006). Automatic Movie Content Analysis: The MoCA Project. http://www.informatik.uni-mannheim.de/pi4/lib/projects/moca/

Prelinger Archives (2006). http://www.archive.org/details/prelinger

Presto (2006). Presto Preservation Technology. http://presto.joanneum.ac.at/index.asp

Price, G. (2006a). Searching for online video. http://searchenginewatch.com/searchday/article.php/3576231

Price, G. (2006b). Searching Television News. http://searchenginewatch.com/searchday/article.php/3582981

Radio Times (2006). http://www.radiotimes.com/

Rapaport, E. (2004). Schoenberg-Hartleben's Pierrot Lunaire: Speech – Poem – Melody – Vocal Perforamnce. *Journal of New Music Research*, *33*, 71–111.

Rayson, P., Leech, G. & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, *2(1)*, 133–152. John Benjamins, Amsterdam/Philadelphia. ISSN 1384-6655. http://www.comp.lancs.ac.uk/ucrel/papers/rlh97.html

Resolume (n.d.). Resolume VJ Software. http://www.resolume.com/features/index.php

Retrieva (2006). Retrieva Overview. http://www.retrieva.com/retrieva/

Rev2.org (2005). Review: Truveo — Impressive Video Search. http://www.rev2.org/archives/2005/11/13/review-truvdeo-impressive-video-search/

Robinson, J. (2006). Personal Communication with H. Nock.

Rosenzweig, R. (2003). Scarcity or Abundance? Preserving the Past in a Digital Era. *The American Historical Review* 108.3. http://www.historycooperative.org/journals/ahr/108.3/rosenzweig.html

Sandom, C. & Enser, P. (2003). Archival moving imagery in the digital environment. In: Anderson, J., Dunning, A. & Fraser, M (eds.) *Digital resources for the humanities 2001–2002*. London: Office for Humanities Communication, King's College. http://www.cmis.brighton.ac.uk/research/vir/DRH2001.pdf

Sapp, C. (n.d.). Reverse Conducting Example. http://mazurka.org.uk/info/revcond/example/

Saraçlar, M., Nock, H. & Khudanpur, S. (2000). Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech and Language*, *14*, 137–160. http://www.busim.ee.boun.edu.tr/%7Emurat/publications/csl00.pdf

Saraçlar, M. & Khudanpur, S. (2004) Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech and Language*, *18(4)*, 375–395. Preprint: http://www.busim.ee.boun.edu.tr/~murat/publications/csl04.pdf

Scansoft (2006). http://www.scansoft.com/audiomining/

Seltzer, W. (2005). Classical Myopia and the BBC's Beethoven. http://copyfight.corante.com/archives/2005/07/12/classical_myopia_and_the_bbcs_beethoven.php

Simons, G. & Bird, S. (2003). The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, *18*, 117–128. http://arxiv.org/ftp/cs/papers/0306/0306040.pdf

Sjölander, K. & Beskow, J. (2000). WaveSurfer – an open source speech tool. *Proc of ICSLP*, Beijing, Oct 16–20, 4:464–467. http://www.speech.kth.se/wavesurfer/wsurf_icslp00.pdf

Sjölander, K. & Beskow, J. (2006). Wavesurfer. http://www.speech.kth.se/wavesurfer/

Sphinx (1996–2004). Sphinx-4: A speech recognizer written entirely in the Java™ programming language. http://cmusphinx.sourceforge.net/sphinx4/

St George's, Leeds (2006). Sermons. http://www.stgeorgesleeds.org.uk/church/sermons.htm

SWAG, Spoken Word Archive Group (2003). Report of the EU/US working group on Spoken Word Digital Audio. http://www.dcs.shef.ac.uk/spandh/projects/swag/swagReport.pdf

Tanghe, K., Lesaffre, M., Degroeve, S., Leman, M., De Baets, B., & Martens, J.-P. (2005) Collecting ground truth annotations for drum detection in polyphonic music. *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 50–57. http://ismir2005.ismir.net/proceedings/1006.pdf

The MathWorks (1994–2006). MATLAB®. http://www.mathworks.com/products/matlab/

Transcriber (2006). A tool for segmenting, labeling and transcribing speech. http://sourceforge.net/projects/trans/

Tranter, S. & Reynolds, D. (2006). An Overview of Automatic Speaker Diarisation Systems. To appear in *IEEE Transactions on Speech and Audio Processing*.

Tucker, S. & Whittaker, S. (2005). Novel techniques for time-compressing speech: an exploratory study. *International Conference on Acoustics, Speech and Signal Proceessing*, Philadelphia. http://www.dcs.shef.ac.uk/~sat/downloads/ICASSP2005.pdf

TV Eyes (2003). TVEyes can monitor television news. http://www.tveyes.com/about_us/news_articles/conpost.htm

Tzanetakis G. (n.d.). MARSYAS: Music Analysis, Retrieval and Synthesis for Audio Signals. http://opihi.cs.uvic.ca/marsyas/

Tzanetakis G. & Cook P. (in press). Audio information retrieval using MARSYAS. In D. Byrd & J.S. Downie (eds.), *Current Research in Music Information Retrieval: Searching Audio, MIDI and Notation* (Kluwer Academic Publishers).

UCLA Film & Television Archive (n.d.). Digital Hitchcock, http://www.cinema.ucla.edu/education/dighitch.html

UCSB, University of California Santa Barbara (2006). Cylinder Preservation Project. http://cylinders.library.ucsb.edu

University of Oxford (2005). British national corpus. http://www.natcorp.ox.ac.uk/

van Leeuwen, D., Martin, A., Przymocki, M. & Bouten, J. (2006). NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech and Language*, *20*, 128–158.

Viddler (2006). Create, enhance and share your video. http://viddler.com/

Virage (2006). Virage Products Overview. http://www.virage.com/content/products/

Volkmer, T. (2006). Efficient Video Annotation (EVA) system. http://domino.research.ibm.com/comm/research.nsf/pages/r.multimedia.innovation.html?Open&printable

von Ahn, L. & Dabbish, L. (2004). Labeling Images with a Computer Game. http://www.cs.cmu.edu/~biglou/ESP.pdf

WCER, Wisconsin Center for Education Research, University of Wisconsin (2006). Transana. http://www.transana.org/

Webb, N., Hepple, M. & Wilks, Y. (2005). Dialogue act classification based on intra-utterance features. *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburg. http://www.dcs.shef.ac.uk/~yorick/papers/AAAI05_A.pdf

Witmore, C. (2005). Multiple fields and archaeological practice: auditory archaeology or the "belles noiseuses". http://metamedia.stanford.edu:3455/multiplefields/1067.

Wolf, W. & Liang, Y. (1997). A Digital Video Library on the World Wide Web. http://portal.acm.org/citation.cfm?id=244457&coll=portal&dl=ACM

Wood, G. & O'Keefe, S. (2005). On Techniques for Content-Based Visual Annotation to Aid Intra-Track Music Navigation. *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 58–65. http://ismir2005.ismir.net/proceedings/1023.pdf

Wright, H., Poesio, M. & Isard, S. (1999). Using high level dialogue information for dialogue act recognition using prosodic features. *Proceedings of an ESCA Tutorial and Research Workshop on Dialogue and Prosody*. http://www.ltg.ed.ac.uk/papers/99wright.pdf

Wrigley, A. (2005a). Archive of Performances of Greek and Roman Drama. http://www.apgrd.ox.ac.uk/

Yahoo! Video (2006). http://video.search.yahoo.com/

Youtube (2006). Broadcast yourself. http://www.youtube.com/

Zhou, B. & Hansen, J.H.L. (2002). SpeechFind: an Experimental On-Line Spoken Document Retrieval System for Historical Audio Archives. *Inter. Conf. on Spoken Language Processing (ICSLP-2002)*, Denver, *3*, 1969–1972. http://cslr.colorado.edu/beginweb/ICSLP2002/p2010.pdf