EU-US WORKING GROUP ON SPOKEN-WORD AUDIO COLLECTIONS

0.0 EXECUTIVE SUMMARY

Our diverse cultures rely increasingly on audio and video resources. We need to chart a steady course to assure the utility of this record. Such a course calls for a plan to preserve these resources and to determine the most effective ways to access their rich content. For example, though our nations possess enormous collections of spoken-word materials, much of these collections will remain inaccessible to the public for lack of adequate search technologies or from decay unless we act to chart an access and preservation path. Our aim is to forge agreement on these vital topics so that as technology changes, we will be able to rely on our collections to understand and preserve these essential components of our cultural heritage. We also need to focus research support on areas of access and preservation that we believe will yield the greatest benefits across many intersecting disciplines. This document presents an agenda for collaborative research in this field.

Spoken-word collections cover many different domains. These include radio and television broadcasts, governmental proceedings, lectures, oral narratives, meetings and telephone conversations. Needs vary in collecting, accessing and preserving such data:

>> Political and economic: providing access for citizens to governmental proceedings, corporate shareholder meetings, public meetings of political parties and NGOs

>> Cultural: building and providing access to large, multilingual archives of broadcast material, public performance, oral narrative

>> Educational: acquisition, preservation, search and access of lectures; use of digital audio and video resources as primary sources for inquiry and explication.

We have now reached the point where various enabling technologies have matured sufficiently for the research community to address these needs.

0.1 RESEARCH AGENDA

We have structured the research agenda for Spoken-Word Archiving into three main areas: technologies, privacy and copyright, and archiving and access. The main priorities are to advance the state-of-the-art within each area and to foster integration among them. It is clear that each area informs the others.

0.1.2. Technologies

Audio/signal processing: Many spoken-word collections of interest, particularly historical collections, have deteriorating audio, due to media degradation or imperfect analog recording technology. Other audio signal processing challenges arise from multiple overlapping speakers (e.g., meetings), low signal quality due to far-field microphones (e.g., in courtrooms), and effects of other sound sources and room acoustics.

Speech and speaker recognition: Any spoken audio collection raises two immediate questions: (a) What was said? (b) Who said it? Speech and speaker recognition technologies now work to minimally acceptable levels in controlled domains such as broadcast news. However, to achieve substantial improvements will require new tools to address less controlled collections of spoken audio. Without such tools, the costs in labor to access spoken-word collections will be prohibitive. The creation of these tools also enables the hearing-impaired public to access and use these materials.

Language identification: In a multilingual context, automatic language identification is essential. In particular many collections (e.g., meetings at a European level, some oral narratives) feature speakers switching between different languages. We can construct adequate baseline systems based on current knowledge, but issues such as within-utterance language change pose interesting and challenging research problems.

Information extraction: The use of a spoken-word collection can be enhanced by the automatic generation of content annotations. Currently it is possible to automatically identify names and numbers and to provide punctuation. However, it would be advantageous to annotate many other elements--particularly prosodic events--above the word level such as emotion, decision points in meetings, and interaction patterns in a conversation.

Collection level browsing and search: The current state-of-the-art for collection-level browsing and search is based on the application of text retrieval approaches to speech recognizer output. While this has been relatively successful in some domains (e.g., broadcast news), such approaches have clear limitations, and the development of new search-and-browse approaches, beyond simple text retrieval, are required.

Presentation: The final technological research area that we have identified is presentation. Currently this involves little more than playing an audio clip and displaying its transcription. There is an enormous need for research in this area. Several examples come to mind: the construction of audio scenes, presentation of higher-level structure, summarization, and presentation of non-lexical information in speech.

0.1.3. Privacy and copyright policy
A number of policy issues arise when discussing spoken-word collections, and it is impossible to treat the technologies in isolation from these issues.

Privacy: Privacy is a major problem, particularly for some spoken-word collections when individuals do not have an expectation that their statements will be archived, although they have spoken in a public forum such as a company board meeting or a political rally. It may not be possible to offer a comprehensive solution to the privacy problem, particularly for materials where contact with the original collector or subject has long since been lost, but research in this area can accomplish some practical goals. Future collectors must be armed with reasonable policies to obtain clearances and document applicable rights.

Copyright: The impact of copyright varies by collection, and by national jurisdiction. Because the legal terrain here is difficult to understand and is undergoing rapid change, a practical approach for cultural institutions to take may be to implement "acceptable risk" policies. These policies set forth overarching principles of respect for subjects and for the creators' intellectual property rights, but balance them against a need to provide access to important cultural heritage materials. Issues to research include: copyright exemptions (e.g., for educational purposes), classes of works that do not qualify for copyright protection, digitization for preservation and mediated access, and questions collection custodians should pose to determine copyright status and likely consequences of wide availability of digital surrogates.

0.1.4. Archiving and access

Preservation: Open research issues include standards for preservation and development of sustainable digital repositories. Issues that need to be addressed include: funding; automating digitization and metadata capture; migrating and refreshing/augmenting collections. Computerized automated capture and preservation of collections clearly underlies the development of this entire area.

Content structure: This area spans metadata, item structure, annotation, discovery and delivery issues, such as network bandwidth. Metadata vocabularies have been developed, but this area still needs further research, particularly when the archived items have a complex structure. Additionally, metadata needs to be aggregated and services offered on the aggregated collection. Models and tools for annotation are a rapidly evolving research area, particularly in the area of distributed and collaborative annotation.

Media storage: Even with the rapidly declining costs of spinning disks, most preservation-quality audio collections will continue to require supplemental digital storage media for the raw audio files at least into the foreseeable future. Research is needed on various media (CD, DVD) and best practices for storing, checking, and refreshing.

0.2 CONCLUSION

Though we represent diverse disciplines, we see convergence in the domain of spoken-word collections to address new and challenging issues. In advancing an ambitious research agenda, we envision ancillary benefits across many communities of interest: speech and language technology; software architecture; information science and digital libraries; and a set of diverse user communities. Progress requires integration across these areas at the international level. In our judgment, the impact will be substantial. To do any less will risk significant loss to an essential element of our collective heritage.

EU:
Steve Renals, CS, University of Sheffield, UK
Franciska de Jong, CS, University of Twente, Netherlands
Marcello Federico, ITC-IRST, Trento, Italy
Lori Lamel, LIMSI-CNRS, France
Fabrizio Sebastiani, IEI-CNR, Pisa, Italy
Richard Wright, BBC Information+Archives, UK

US:
Jerry Goldman, Political Science, Northwestern University
Steven Bird, CIS/LDC, University of Pennsylvania
Claire Stewart, Library, Northwestern University
Carl Fleischhauer, Library of Congress
Mark Kornbluh, MATRIX and History, Michigan State
Douglas W. Oard, CLIS/UMIACS, University of Maryland

1.0. INTRODUCTION

This report emerges from a joint working group, supported in the US by the NSF Digital Library Initiative and in the EU by the Network of Excellence for Digital Libraries (DELOS). The aim of the group is to define a common agenda for research in the area of spoken word audio collections, and to define areas for collaborative research between European and US researchers. The scope of this report is spoken word audio, with no explicit reference to related areas, such as non-speech audio or video.

The following sections set out the issues, approaches and polices that converge in the area of spoken-word collections. Section 2 outlines the major issues in the field. Section 3 surveys the current technological state-of-the-art. Section 4 examines the controversial and rapidly changing policy issues pertaining to privacy and copyright. Section 5 covers the issues of collecting, archiving and preserving spoken-word content. An appendix addresses content preservation from a digital library perspective.

2.O DESCRIPTION OF THE ISSUES

2.1 Types of spoken word collections

As recording technology improved dramatically over the course of the twentieth century, the size and diversity of spoken word audio collections expanded geometrically. The cost of recording and preserving sound has moved steadily downward as progressive technological change has facilitated successive generations of recording devices, each with increased storage capacity. While some recordings have been migrated forward to newer technologies, archives today contain recorded collections from every stage of technological development, captured at diverse standards, and on assorted storage media.

Existing spoken word collections cover an enormous range, from the earliest recordings of public speeches and broadcasts on wax cylinders and 78 rpm records to oral histories on cassette and reel-to-reel tapes and on through continuous digital recordings of contemporary broadcast news.

With declining costs and increasing recording and storage capacity, the breadth of audio collections continues to grow. As a result, the domains covered by spoken word collections are both varied and vast. Nonetheless, it is possible to identify the major types of spoken word collections. These are:

1. Broadcast news (both radio and TV)
2. Governmental proceedings (parliamentary debates, court recordings, commissions and committees)
3. Presentations in the form of speeches, sermons, and lectures (political, religious, educational)
4. Oral narratives (usually retrospective interviews)
5. Interactive meetings (business (e.g., shareholders), political (e.g., political party conventions)
6. Recorded telephone conversations

2.2 Magnitude of content

National archives and public broadcast archives in Europe and the United States have millions of hours of holdings, much of which features spoken language. The bulk of

this content -- estimated to be 80 percent -- is in analog form. It will perish within a few decades unless we take steps to preserve it. An order-of-magnitude estimate for all significant world holdings of audio material in analog formats is 100 million hours. In addition, millions of hours of spoken language materials come into existence in digital form every year. As digital systems replace analog systems, and as recording and storage costs decline, we envision accelerated growth in the creation of spoken word documents and increased demand for efficient archiving and retrieval strategies.

2.3 User communities

Not surprisingly, the user communities for spoken word collections are as vast and varied as the collections themselves. Indeed, many collections serve diverse needs for very different user communities. For example, a recorded speech might be used for political purposes, for educational and research purposes, or for linguistic analysis.

Spoken word collections have implications across all areas of daily life. In politics, recordings of speeches, debates, broadcasts, etc. are an essential source of governmental proceedings, political candidates and positions, and citizen activities. In commerce, recordings of board meetings, political debates, broadcast news, etc. can provide access to vital information for economic planning and action. In law, recordings can serve as essential evidence in civil and criminal cases. And in culture and education, spoken word collections are vital to preserve, understand and teach about all aspects of social life.

2.4 Archives

Existing spoken word archives are equally varied. Government bodies, archives, libraries, museums, universities, churches, political parties, corporations, broadcasters, recording companies, community organizations, and individuals hold spoken word collections. While some major media organizations such as broadcasters and recording companies have prioritized preservation and developed systems for access, particularly with native digital collections, these are the exceptions. For most organizations, their spoken word collections comprise a small part of a larger effort to collect written material. Small and regional media organizations, which often produce large amounts of new spoken word materials daily, do not have the resources to archive their growing collections adequately, if at all. Thus spoken word collections are often the stepchild of an archive—minimally managed, poorly preserved, and hardly accessible.

Most spoken word archives have analog holdings on various media, typically different types of tape. Tape recordings have limited life spans, are hard to maintain and lose quality when copied on to new analog media. Analog recordings can provide only linear access; they must be listened to in consecutive order. Preservation needs and access are in perpetual conflict with analog materials since the media deteriorate with use. Working or shelf copies protect the source material, but increase the archiving costs through additional physical space and collection management.

The future of spoken word archives clearly lies with digital technology. Most new recordings, including broadcast materials, are now 'born digital.' Equally important, conversion of older analog content to digital media is essential for both long-term preservation and to provide increased access to spoken word resources. With digital collections, access and preservation are not in conflict. Digital content can be endlessly replicated with no loss of quality. Most important, digitization breaks the

linear tyranny of analog recording. Digital sound can be searched in ways unimaginable with analog recordings. Access can be nearly instantaneous.

2.5 Access

For analog recordings, access is only possible through replication of the physical recording. To listen to most spoken word recordings in archival collections today, one must either travel to the archive and listen to a second- or third-generation copy or pay for a copy of the second- or third-generation recording. The cost in time and resources constrains access. Digital recordings, however, can be transmitted without loss over the Internet. The World Wide Web is rapidly becoming a doorway to digital audio collections. Streamed audio access to news and cultural programming can be accessed from RAI Radio (Italy) <http://www.radio.rai.it/>, BBCi (United Kingdom) <www.bbc.co.uk>, and National Public Radio (United States) <www.npr.org>. Other specialized spoken word collections, such as The OYEZ Project <www.oyez.org>,which delivers the archived recordings of US Supreme Court arguments <www.oyez.org>, vividly illustrate the increased access to spoken word collections made possible by advances in information technology. The Web, however, is only the first doorway to digital spoken word collections. Already we are seeing the development of alternative multimedia delivery devices, from collection sharing via peer-to-peer networking and downloading of spoken-word materials in popular formats like MP3 or OGG. Client storage devices hold ever-increasing amounts of data at lower and lower price points. PDAs and cell phones coupled with the development of new services will soon permit multi-channel delivery including spoken word materials such as talking books.

2.6 Convergence of Technology, Needs, and Possibilities

We have now reached the point where various enabling technologies have matured sufficiently for large-scale conversion of spoken word resources from analog to digital. Computer equipment is now available enabling low cost and low loss transfer to digital media. The archival community has set standards for such conversion to ensure the integrity of the original collections. Ubiquitous and inexpensive computers offer ready access to digitized sound.

These are only the first steps, however, to ensuring long term preservation and enhanced access to spoken word collections. We are at a watershed moment where digital sound archives are possible, their advantages over old analog collections are self-evident, and the technology to work with digital audio has matured to facilitate a new research agenda for spoken word archives. Furthermore, many of these analog collections are at risk, and we have a narrow window of opportunity to preserve digitally this analog media, before it starts to become unusable. We can now envision with confidence and clarity the research agenda in speech recognition, speech enhancement, speech analysis and retrieval, archival design, metadata development, delivery interface, educational integration, and other areas to fully realize the potential of spoken word archives.

The digital revolution has the potential to do for aural resources what the printing press did for written resources. For the first time in human history, the spoken word can now be preserved for the long term and made accessible to those far beyond hearing range and in ways that open up entirely new possibilities for human culture. Researchers have the tools and capacities to transform access to the spoken word and vastly enrich capacities across all aspects of our societies.

3.0 CURRENT TECHNOLOGIES AND NEW LANDSCAPES

In the last decade, speech recognition technology has made impressive advances and has proven to be effective for indexing audiovisual archives. Research projects, such as Informedia at Carnegie Mellon University <http://www.informedia.cs.cmu.edu/>, and recent commercial products, such as Virage <http://www.virage.com/>, have successfully deployed state-of-the-art speech recognition into digital libraries of broadcast news. Automatic indexing and content-based access of audiovisual archives is today feasible thanks to outstanding results from research in speech recognition, language processing, and information retrieval. An important driving force in speech recognition has been the US Defense Advanced Research Projects Agency (DARPA)<www.darpa.mil>. Working through the National Institute of Standards and Technology (NIST)<www.nist.org/speech> and the Linguistic Data Consortium at the University of Pennsylvania (LDC) <www.ldc.upenn.edu>, DARPA coordinates and focuses efforts on relevant research topics, collects language resources, and organizes systematic evaluations. We survey technologies for indexing and accessing spoken documents developed under the DARPA umbrella.

3.1. Background Information.

Work in Large Vocabulary Continuous Speech Recognition (LVCSR) started well over a decade ago with tasks mostly oriented toward automatic dictation. In the United States, DARPA set up a common framework by providing both large amounts of training data and evaluation data. The reference task--dictation of Wall Street Journal articles--was used to refine the technology for dealing with a vocabulary of several tens of thousands of words, a challenging task in itself at the time. In the years that followed, research interest moved toward making automatic speech recognition (ASR) systems capable of handling a range of acoustic conditions and speaking styles much wider than what could be found in dictation tasks. The new reference task became therefore the transcription of broadcast news (BN) programs, for which increasing amounts of training data were collected.

During the same years, similar evaluations were organized by DARPA for other research problems related with spoken information processing: speaker recognition, information extraction, spoken document retrieval, and topic detection and tracking. Tasks of increasing complexity have been defined over time, for which increasing amounts of training data have been made available to participants. As a consequence, scaling-up of the technology was enforced, together with improvements in performance and robustness of the systems. Concerning the content of data, up to now most of the evaluations have been carried out on American English BN. The news domain is indeed very general and makes data collection relatively easy. Research aimed at porting these techniques to other domains and languages has started in several labs.

3.2 Audio Indexing

Audio indexing involves several discrete topics: audio partitioning, speech recognition, speaker identification, information extraction, and automatic summarization. We examine each topic below.

3.2.1. *Audio Partitioning*

Audio partitioning is concerned with segmenting an audio stream into acoustically homogeneous chunks and classifying them according to a set of broad acoustic classes. For instance, for the purpose of speech recognition, the audio is usually partitioned by identifying segments containing speech versus other types of content,

such as music. In many systems, the classification of speech segments is refined by recognizing, for instance, the signal bandwidth, the gender of the speaker, the speaker itself, the level of noise, etc. The difficulty of this task increases with the level of detail required by the segmentation/classification task. For instance, while detecting speech segments in conversational speech is relatively easy, detecting speaker turns can be very difficult when overlapping speech occurs (that is, when two people speak simultaneously). Moreover, segment classification, as well as any other pattern classification task, becomes difficult when the actual conditions mismatch with those observed in the training data. Acoustic segmentation and classification is crucial for indexing audio recording which may contain more than pure such, e.g. music scores, jingles, etc. Moreover, an accurate segmentation can be exploited to run speech recognizers specifically trained on a given acoustic condition, e.g. bandwidth, gender, speaker.

In recent years, several algorithms have been presented which use a statistical decision criterion to detect spectral changes (SCs) within the feature space of the signal. Assuming that a Gaussian process generates data, SCs are detected within a sliding window through a model selection method. The most likely SC is tested by comparing two hypotheses: (i) the data in the window are generated by the same distribution; (ii) the data in the left and right halves of the window are drawn by two different distributions. The test is performed with a likelihood ratio that also takes into account the different "sizes" of the compared models. Usually, the Bayesian Information Criterion (BIC) is applied to select the best fitting model.

In order to classify segments, researchers use Gaussian mixture models, which typically have been trained on supervised data. Finally, clustering of speech segments is carried out by a bottom-up scheme that groups segments, which are acoustically close with respect to the BIC or some defined metric. Audio partitioning has been applied successfully and extensively, mainly on broadcast news transcription. The application to other audio collections poses problems of portability and robustness of the methods, which at the moment are surmounted by tuning the system on some development data. Future work should be devoted to developing robust methods which can cope with greater variability of acoustic conditions.

3.2.3. *Speech Enhancement*

Speech is often recorded under sub-optimal conditions, but pre-processing techniques can be used to enhance the suitability of the signal for subsequent processing. For access to spoken content, speech enhancement typically seeks to achieve one or more of the following goals: (1) improved accuracy from subsequent automatic processing (e.g., automatic speech recognition), (2) improved intelligibility for a human listener, or (3) a qualitative improvement in the listening experience for a human listener. Human perception is far more robust than present automated approaches to speech recognition, so signal processing that precedes speech recognition is presently the focus of a substantial research effort. The initial focus of that work has been accommodation of environmental factors (e.g., background sounds such as vehicle noise or unrelated transient signals, and the results of microphone placement and room acoustics such as echo or reverberation) and the effect of transmission channels (e.g., speech compression algorithms for cellular telephones). Work with recorded materials has generally focused on improving intelligibility and/or the listening experience, topics often referred to as "audio restoration." Much recorded speech is stored on analog media, including cassette tape, open-reel magnetic tape, phonograph records, and (less commonly) Dictabelt loops, wire recordings and wax cylinders. In addition to environmental factors,

analog recordings might be degraded when they are first created (e.g., by the frequency response of the microphone), during duplication (e.g., reduction in the signal-to-noise ratio), during storage (e.g., warping of a phonograph record), as a result of prior use (e.g., splicing to repair a tape break), and during replay (e.g., due to variations in motor speed). Audio restoration techniques leverage an understanding of the characteristics of undesirable signal components (e.g., clicks and pops from damaged phonograph records, or "thumps" from Dictabelt loops that have were folded for storage) and human perceptual characteristics (e.g., critical bands and auditory masking) to produce a more satisfactory reproduction of the original content.

3.2.4. *Speech recognition*

Speech recognition is concerned with converting the speech waveform (an acoustic signal) into a sequence of words. In the context of audio archives, the audio signal often contains more than just speech. These other sounds may be intentionally recorded: such as background music or noise added to set the mood, or samples of sounds such as animal vocalizations.

Speech is generally produced with the purpose of being understood by a native speaker of the same language, who usually shares some set of common values or experience with the speaker. The choice of lexical items and speaking style depend on the given talker and the intended audience. There are significant differences of an acoustic nature due to anatomical differences across speakers, as well as social and dialectical conventions. These factors complicate speech understanding for humans and machines. Transcribing and annotating audio data are necessary to provide access to its content, and large vocabulary continuous speech recognition is a key technology for automatic processing. Such audio data is challenging to process as it consists of a continuous flow of audio data comprised of segments with various acoustic and linguistic characteristics. Processing such inhomogeneous data thus requires appropriate modeling at the acoustic and linguistic levels. (Since much of the linguistic information is encoded in the audio channel of video data, once transcribed it can be accessed using text-based tools. Transcripts will allow users to access data based on linguistic content.)

Today's most effective speech recognition approaches are based on a statistical model of the speech signal. Speech is assumed to be generated by a language model which provides estimates of the probability of all word strings independently of the observed signal, and an acoustic model encoding the message in the audio signal. The goal of speech recognition is to find the most likely word sequence given the observed acoustic signal.

Transcription system development requires large annotated training corpora for all languages and audio data types of interest. Transcription performance is highly dependent upon the availability of sufficient training materials, the preparation of which requires substantial human effort. Speaker independence is obtained by estimating the parameters of the acoustic models on large speech corpora containing data from a large speaker population. It is common practice to use gender-dependent acoustic models to account for anatomical differences (on average, females have a shorter vocal tract) and social ones (female voice is often "breathier" caused by incomplete closure of the vocal folds). Other groupings of speakers according to different characteristics such as dialect or speaking rate have also been investigated to improve performance. State-of-the-art systems are typically trained on several tens to hundreds of hours of audio data and several hundred million

words of text materials. The significant advances in speech recognition over the last decade can be partially attributed to advances in robust feature extraction, acoustic modeling with effective parameter sharing, unsupervised adaptation to speaker and environmental condition, efficient dynamic network decoding, and audio stream partitioning algorithms, as well as to the availability of large audio and text corpora for model estimation, combined with increased computational power and storage capacity.

While the same basic transcription technology has been successfully applied to different languages and types of speech, specific adaptations are required to optimize performance. Mismatches in training and test conditions typically result in high error rates.

Despite significant advances in speech recognition, at least two fundamental problems remain: speed and robustness. There is a large gap between machine and human performance (a factor 5 to 10, depending upon the transcription task and test conditions). It is well acknowledged that there are large performance differences for the best systems (attributed to a variety of factors such as speaking style, speaking rate and accent). Improvements are needed in the modeling techniques at all levels: acoustic, lexical and pronunciation, and linguistic (syntactic and semantic).

Ongoing research [4] is addressing issues such as reducing the cost of system development [18], and improving the genericity, portability [19, 2] and adaptability of the models. Some techniques of interest are, for example, light and unsupervised training, faster adaptation techniques, learnable pronunciation lexicons, language model adaptation, topic detection and labeling, and metadata annotation. Accurate metadata annotation (topic, speaker, acoustic conditions) can also be used to adapt generic models to the particular audio data type to be transcribed.

3.2.5. *Speaker identification and tracking*

Speaker recognition has been an active research area for many years [32, 7]. Several types of recognition problems can be distinguished: speaker identification, speaker detection and tracking, speaker verification (also called speaker authentication). In speaker identification the absolute identity of the talker is determined. In contrast, for speaker verification the question is to determine if the talker is the person s/he claims to be. Speaker tracking refers to finding audio segments from the same speaker, even if the identity of the speaker is unknown.

Accurately identifying a speaker [23, 14] is an unsolved research problem, despite several decades of research. The problem is quite close to that of speech recognition in that the speech signal encodes both linguistic information (i.e. the word sequence which is of interest for speech recognition) and non-linguistic information (the speaker identity, as well as less well-quantified values such as mood, emotion, attention level, etc.). The characteristics of a given individual's voice change over time (short and long periods) and depend on the talker's emotional and physical state. The identification problem is also highly influenced by the environmental, recording, and channel conditions. For example, it is very difficult to determine if a voice is the same in different background conditions, such as in the presence of background music or noise.

Automatically identifying speakers and tracking them throughout individual recordings and in recording collections can reduce the manual effort required to

annotate this type of metadata. Automatic speaker identification will allow digital library users to access spoken word documents based on who is talking. Some of the recent speaker tracking research can potentially allow talkers to be located in large audio corpora using a sample of speech, even if the absolute identify of the talker is unknown.

Most of today's working speaker recognition systems [33, 30] make use of the same statistical approaches as are used in speech recognition, i.e., hidden Markov models or Gaussian mixture models. Speaker specific models estimated on speaker-specific audio data are used to assess whether unknown speech samples are from one of the known speakers (speaker identification). Much of the research in speaker recognition has been for security purposes, either controlling access to a physical location or to restricted information, or in intelligence monitoring. Recent promising research at the Johns Hopkins Summer Workshop 2002 (SuperSID: Exploiting High Level Information for High-performance Speaker Recognition) <http://www.clsp.jhu.edi/ws2002> has addressed using multiple types of acoustic, supra-linguistic and phonetic attributes to improve speaker recognition performance.

While today's speaker recognition technology is not perfect, performance levels are probably adequate for use in automatic annotation of audio collections and for speaker-based access in digital libraries.

3.2.6. *Information extraction*
Information extraction (IE) is the task of extracting meaningful information from information sources. The search objective can range from named entities -- such as persons, organizations, and locations -- to attributes, facts, or events. The difficulty of information extraction is related to the natural language processing required to recognize complex concepts, the intrinsic ambiguity of named entities (e.g., the name "Barcelona" could denote a city or a football team, depending on the context), and the steady evolution of language (e.g., new names gradually appear in the media).

Given the aim of accessing spoken documents, information extraction automatically selects pieces of content that may prove interesting or useful. Moreover, by maintaining links between the extracted information and the original documents, it is possible to provide context for each retrieved concept. Most recent research on information extraction from spoken documents has been carried out under the IE Entity Recognition and Automatic Content Extraction (ACE) programs under DARPA and NIST. Considered tasks are the detection of named entities (names of locations, organizations, and people), temporal expressions, currency amounts, and percentages, within BN shows. State-of-the-art performance was achieved as well by rule-based system and statistical language modeling approaches. Research under the ACE program currently focuses on more complex tasks, such as detecting and tracking entities over time, recognizing mentioned events, and relations among entities.

3.2.7. *Automatic summarization*
Speech summarization is commonly applied to techniques that reduce the size of automatically generated transcripts in a way that resembles summarization technology for text documents. Its goal can thus be described in a similar way as text summarization: to take a partial or unstructured source text, extract information content from it, and present the most important content in a condensed form in a manner sensitive to the needs of the user and task. Depending on the nature of the

content and the user information need, both summarization of single fragments as well as multi-document summarization can be helpful browsing tools.

The fact that speech transcripts may be linguistically incorrect requires techniques for enhancement of the content. To generate coherent, syntactically well-formed descriptions that preserve the original meaning, semantically complex operations have to be developed, e.g., for anaphora resolution. Two types of summarization tasks are distinguished here: (1) condensation of content to reduce the size of a transcription according to a target compression ratio, e.g. to produce closed captions, meeting minutes, etc, involve both intra-sentential as well as text processing, and (2) a presentation tool for spoken document retrieval. Several obstacles impede transparent presentation of speech retrieval results. Automatically generated audio transcripts are not easily read, because of recognition errors and the lack of punctuation, but also because of disfluencies, repairs, repetitions, etc. Extraction of the relatively important information can help users to browse more easily through search results.

Purely audio summaries of speech can be envisaged, and prototype speech skimming systems have been developed. An important issue in this case is the development of accelerated audio playback, which is an interesting signal-processing task if intelligibility and the speech characteristics (such as intonation) are to be maintained as much as possible. This area is rather closely related to speech synthesis.

3.2.8. *Prosody*
A spoken message contains more that simply what was said (i.e., the text transcription) and who said it (i.e., the identity of the speaker). The prosody (timing, intonation and stress) of the speech signal offers a great deal more information such as the emotional state of the speaker, boundaries and "punctuation" in the speech and disambiguation of the intended message (e,g., questions have a rising intonation). The research challenge in this area is to develop prosodic models that are sensitive to these supra-segmental features.

3.3 Collection Level Browsing and Searching
Collection level browsing and searching involves several complex tasks including: spoken document retrieval, interactive speech retrieval, topic detection and tracking, cross-language information retrieval, and speech enhancement. We take up these topics below.

3.3.1. *Spoken document retrieval (SDR)*
By using speech recognition to convert speech into text, detailed textual representations can be generated for spoken content. These representations are not exact renderings of the spoken content, but they do allow searching for specific words and phrases and in general are suited for a variety of audio browsing support tools. Since speech recognition systems can label recognized words with exact time stamps, their output can be viewed as metadata by which it becomes possible to lead users directly to relevant audio fragments (perhaps with links to related content, e.g., video). By default, SDR recognition technology is speaker-independent and geared toward continuous speech and large vocabularies. Building an acoustic and language model requires substantial effort and data. For general-purpose audio mining tools, acceptable retrieval performance calls for a minimum word error rate of .50.

Tuning the lexica to specific domains, collections or periods require additional effort and work flow procedures from user organizations. Recognition of unknown words (numbers for compounding languages like German and Dutch are relatively high) and proper names are problematic. Many audio collections are difficult to search because of the recording conditions (e.g., multiple speakers, bandwidth, background noise) do not meet minimum requirements. Transparent presentation of retrieval results is hindered in several ways. It is not easy to ready audio transcripts due to these errors and the lack of inter-punctuation. Simply put, retrieval requires listening. But semantically sound fragment boundaries are not easy to detect, complicating listening retrieval. Therefore fragment and cluster classification is crucial to SDR.

SDR allows the disclosure of speech at the fragment level in a way that resembles the most common text search engines. Other search support techniques, e.g., automatic classification and clustering, are applicable on automatic transcripts. We distinguish two approaches: (1) word-spotting and (2) automatic transcript generation in combination with (advanced) full text retrieval tools. For word-spotting, acoustic models are built for a small set of words that are matched during retrieval on query-term models. However, word-spotting is only suited for a small set of search terms. Its chief advantage is that is requires no off-line content processing. Automatic transcription requires acoustic models, (statistical) language models (co-occurrence frequencies) and a recognition lexicon (for some systems limited to 65k words). Its principal limitations are the requirement of off-line content processing and the availability of large corpora. A lot is uncertain about the retrieval performance for speech content. Commercial audio mining tools are available for English only. Systems have been compared only within the general news domain. Within DARPA context, speech retrieval is considered a solved problem. [13]. However this is only valid in a very academic interpretation of the concept of SDR.

Many open problems remain and we envision substantial issues calling for additional research. There is little experience with SDR outside research labs. Content segmentation at a semantic level is crucial, but poorly developed. Current technologies for recognition require huge textual training collections and labour-intensive annotation of audio training corpora. These investments are not straightforward for smaller languages. Techniques for training that circumvent the annotation task are under investigation. Evaluation measures specific for SDR recognition technology are not generally available. (Word error rate is not always the best predictor of retrieval quality.)

3.3.2. *Interactive Speech Retrieval*

Any interactive search process involves five stages, as represented in Figure 1. In query formulation, users interact with the system to craft an expression of the information need--a query--that the system can use to produce a useful search result. Queries are typically expressed as either an undifferentiated set of search terms or as a Boolean expression. In the sorting stage, the system reorders the documents, seeking to put the most promising recordings ahead of others. In Boolean systems, this typically equates to placing documents into one of two sets (relevant, or not). Increasingly common "ranked retrieval" systems take a different approach, allowing searchers to pose queries with little or no structure and then peruse a ranked list of potentially interesting recordings.

Efficient indexing enables quick searches of large collections, but the effectiveness of interactive searching ultimately depends on synergy with a sophisticated user.

Humans bring sophisticated pattern recognition, abstraction and inferential skills to the search process, but the number of documents to which those skills can usefully be applied is limited. The goal of the selection stage is to allow the user to rapidly discover the most promising documents from a system-ranked list through examination of indicative summaries, i.e., summaries designed to support selection. These summaries are generally quite terse since several must be displayed simultaneously in the available screen space. Because summaries may not provide enough information to support a final selection decision, modern systems also typically provide users with the ability to play segments of individual recordings. Direct use of a recording may also result from replay within the retrieval system, or a separate delivery stage may be required (e.g., the audio might be stored on a compact disk for later replay with high fidelity).

Recorded speech poses both challenges and opportunities for the interactive retrieval process. The key challenges are deceptively simple: automatic transcription is imperfect and listening to recordings can be time consuming. Some important opportunities include potential use of speaker identification, speaker turn detection, dialog structure, channel characteristics (e.g., telephone vs. recording studio) and associated audio (e.g., background sounds) to enhance either the sorting or the browsing process. Multimedia integration (e.g., with video) also offers some important opportunities for synergy. For example, query formulation based on spoken words might be coupled with selection based on key frames extracted from video.

3.3.3. *Topic Detection and Tracking*

The Text Retrieval Conference's (TREC) Spoken Document Retrieval (SDR) track emerged in 1996 from a tradition of ranked retrieval evaluations, and the design of the track reflects that heritage. In 1997, a second venue for comparative evaluation of speech retrieval research was introduced in the United States; it is known as Topic Detection and Tracking (TDT). The still ongoing TDT evaluations reflect a broadening of speech processing research to include a strong application focus. Four test collections (known as TDT-1 through TDT-4) have been developed, with the most recent having the following distinguishing characteristics: (1) multi-modal, including both broadcast news audio and newswire text; (2) multilingual, including English, Chinese, and Arabic; (3) multi-source, typically including news from more than one source in each combination of modality and language; (4) event-oriented, with relevance assessment based on whether a story reports on a relatively narrowly defined event (e.g., a specific airplane crash).

The most recent TDT evaluations include comparative evaluations on five tasks: (1) topic segmentation, in which systems seek to discern the times at which the story being reported changes; (2) topic detection, an unsupervised learning task in which systems seek to cluster stories together if they report on the same event; (3) topic tracking, a semi-supervised learning task in which systems seek to identify subsequent news stories that report on the same event as one or more example stories; (4) new event detection, in which systems seek to identify the first story to report on each event; and (5) story link detection, in which systems seek to determine whether pairs of stories report on the same event. The story segmentation task is performed only on broadcast news sources. All other tasks are multi-modal.

3.3.4. *Cross-Language Information Retrieval*

When searchers lack the language skills needed to pose their query using terms from the same language as the spoken content that they seek, some form of support for

translation must be embedded within the search system. There are three cases in which such a capability might be useful: (1) use by searchers capable of understanding the spoken language who are not sufficiently fluent to formulate effective queries in that language (e.g., searchers with a limited active vocabulary in that language); (2) use by searchers lacking the ability to understand the spoken language, if their principal goal is to find easily recognized objects associated with the spoken content (e.g., searching photographs based on spoken captions); an (3) use by any searcher, if suitable speech-to-speech (or speech-to-text) translation technology can be provided. At present, speech-to-speech translation has been demonstrated only in limited domains (e.g., travel planning), but development of more advanced capabilities are the focus of a substantial research investment.

Cross-language information retrieval relies on three commonly used strategies: (1) query translation, (2) document translation, and (3) interlingual techniques. Query-translation architectures are well suited to situations where many query languages must be supported. In interactive applications, query translation also offers the possibility of exploiting interaction designs that might help the system better understand the system's capabilities and/or help the system better translate the searcher's intended meaning. "Document translation" is actually somewhat of a misnomer, since it is the internal representation of the spoken content that is translated. Document translation architectures are well suited to cases in which query-time efficiency is an important concern. Document translation also typically offers a greater range of possibilities for exploiting linguistic knowledge because spoken content typically contains many more words than a query, and because queries are often not grammatically well formed. With interlingual techniques, both the query and the document representations are transformed into some third representation to facilitate comparisons. Interlingual techniques may be preferred in cases where many query languages and many document languages must be accommodated simultaneously, or in cases where the conforming space is automatically constructed based on statistical analysis of texts in each language.

4.0 PRIVACY AND COPYRIGHT

Collectors of spoken word audio materials must address a number of complex privacy and copyright issues relating to the collection, retention and distribution of works. These policy issues cannot be ignored, but the legal frameworks that define them offer incomplete and sometimes conflicting guidance. Privacy and copyright are two of the most rapidly changing aspects of United States and European law. We provide a brief analysis of some key issues.

4.1 Privacy

Privacy is not a precisely defined concept. The issues of data and communications privacy have been very widely debated, across both the U.S. and the E.U. Less commonly discussed aspects of privacy may be equally relevant to a spoken-word archive. For example, what are the legal implications of recording the proceedings of a public meeting?

4.1.1. *An expectation of privacy*

Some issues surrounding audio and video capture in public are not dissimilar to those debated when face-recognition technology began to be used to scan for potential criminals in crowds at airports and other public places.[25] Here, the expectation of privacy is one of anonymity, but this expectation is not always codified in law. Several U.S. state courts have resisted attempts to curtail video and

audio recording in public, finding that no reasonable expectation of privacy can exist in a public place.[34] Use of recording technologies for public surveillance in the United Kingdom has been common for some years, though the government in 2000 signaled its intention to regulate such surveillance in accordance with its 1998 Data Protection Act, passed to harmonize U.K. laws with the 1995 European Union Data Protection Directive.[16] Other E.U. nations, including Greece and Sweden, also interpret the E.U. Directive (revised in 1998 and 2000) to specifically pertain to public video surveillance and closely regulate its use. [25]

Use of wiretapping and other communications surveillance technology is, in general, well regulated, requiring that law enforcement obtain court or judicial orders to make use of such know-how. In reality, permission to wiretap is easily obtained. In the United States no state or federal law enforcement agency requests for wiretaps were denied in 2001, and a total of 1,491 were authorized.[12] The French government approved 4,175 wiretaps in 2000, and the German government 12,651 in the same year.[25: pages 178, 185, 388] Open monitoring and recording of telephone transactions and monitoring of employees' electronic communications for business purposes is also widespread.[9] The right of employees to opt out of such data-gathering has been weak or non-existent. The E.U. is leading the push to expand data privacy regulations to include employee-monitoring activities, which may have the effect of discouraging such monitoring beyond the E.U.[15] Most European Union nations have appointed a central data protection agency, charged with oversight of all personal data collection and processing, and grant individual citizens a mechanism for review, change or removal of their own information.

The ability of governments to compel disclosure of recordings, data, and personal information has increased since 2001, particularly for electronic communications, and particularly in the United States. Under the October 2001 USA PATRIOT Act, federal law enforcement agencies are still required to obtain permission to access records; but the agencies now have the additional instrument of a Foreign Intelligence Surveillance Act (FISA) order along with warrants and subpoenas. FISA, passed in 1978, created a secret court that acts on terrorism investigations in national security situations.[36]

Given the need for oversight and the ease of access to such information once stored in digital form, some difficult choices face the custodian. What balance should be struck between protection of the individual and benefits of large spoken word collections for worthy public purposes (e.g., scholarly inquiry, political discourse, law enforcement, artistic expression)? A good place to turn for examples and guidance may be the regulations governing research on human subjects.[9] These regulations clearly advocate informed consent and limited gathering and use of personal data.[42]

Collecting agencies should determine whether individuals have granted permission for a recording to be made, implicitly or explicitly. A signed consent or permission form is the best safeguard, but is unlikely to be available, particularly for older recordings. Presenters and announcers, interviewers and interviewees, audience members and call-in guests, parties in a conversation: all such participants must be considered when determining whether privacy rights are an issue. A public figure, such as a politician or a known lecturer, is unlikely to substantiate an invasion of privacy claim were his speech to be recorded. The more public the citizen, the less likely he or she is to be able to make a claim.

4.2 Copyright
There are three main issues concerning spoken word materials:

1. whether these materials are protected by copyright,
2. whether rights auxiliary to the copyright must be taken into consideration when considering an archiving digitization initiative, and
3. all rights notwithstanding, whether an argument can be made to proceed with digitization and delivery.

Copyright legislation has changed dramatically over the past decade, both in the United States and in Europe. The rise and demise of Napster and other online file-swapping services have focused the attention of the technology, content, legal and consumer advocacy communities on the issue of digital audio distribution. Despite this attention and debate, clear rules have failed to emerge, and are unlikely to surface in the near term, particularly for non-commercial use by libraries and archives.

4.2.1. *Extent of copyright protections for spoken word materials*
As signatories to the Berne convention [43], the United States and the European Union member nations have reciprocity in copyright protection so that materials created or published in one nation will, for the most part, enjoy the same protections in other nations. Copyright statutes generally reserve for the copyright holder the exclusive right to reproduce, display, distribute copies of, and perform or broadcast the work. The European Union issued a copyright directive [11] in 2001 that matches many of the provisions in the United States Digital Millennium Copyright Act (DMCA) of 1998. Both extend encryption protections with harsh anti-circumvention language. Principles in the EU Copyright Directive will be implemented through the laws of member nations. The results of this implementation do not yet offer clarity or guidance.

In general, sound recordings have historically been accorded fewer protections than other types of works, though some recent initiatives have the effect of increasing theirprotection.[17] In the United States, sound recordings were not protected by federal copyright law until 1972, and recordings made before that date are still not federally protected (though they may be under state copyright laws). Works fixed after 1977 receive at least 70 years of protection. (In the United States, in order for works to qualify for protection, they must be fixed in some physical medium. This requirement has been clarified to encompass digital publication, as well.) In the United Kingdom, copyright for sound recordings was established in the 1911 law [24] and lasts for 50 years, 20 fewer years than granted to creators of print works. The 1979 revision of the Berne convention likewise established a 50-year duration of copyright, a term also endorsed by the European Union in 1993.[5]

Most of the signatory nations require either some form of fixity (United States) or availability to the public (Germany and the United Kingdom) in order to claim copyright protection. However, France's copyright law is much more generous toward authors, stating: "A work shall be deemed to have been created, irrespective of any public disclosure, by the mere fact of realization of the author’s concept, even if incomplete."[28]

There may be layers of authorship embedded in a single sound recording, and each act of authorship may be subject to separate protection. For a musical work, the composition and arrangement might both be protected even if the physical recording

itself is not. A more relevant example of layered rights may be seen in observing several separate acts of creation that might be said to be encompassed within a sound recording of a news broadcast: a typescript, background music, and interviews with news subjects. It is unclear how stringently these protections will be pursued and enforced.

The Berne convention singles out certain types of works and suggests that signatory states may wish to exempt them from copyright protection. The article reads in part:

> (1) It shall be a matter for legislation in the countries of the Union to exclude, wholly or in part, from the protection provided by the preceding Article political speeches and speeches delivered in the course of legal proceedings.
>
> (2) It shall also be a matter for legislation in the countries of the Union to determine the conditions under which lectures, addresses and other works of the same nature which are delivered in public may be reproduced by the press, broadcast, communicated to the public by wire and made the subject of public communication as envisaged in Article 11bis(1) of this Convention, when such use is justified by the informatory purpose.[1]

One possible interpretation of the article is that, in some countries, recordings of lectures, speeches and courtroom oral arguments, and other such public speech enjoy fewer protections under the law than other forms of expression.

4.3 Moral rights

In addition to the set of rights recognized in copyright laws, other rights may come into question with audio archiving projects. Among these are so-called "moral rights," those that allow the creator of a work some lasting ability to control the context in which their works are used and how (or whether) authorship is attributed. Generally, copyright governs economic rights, but moral rights are less tangible and involve the integrity of a work.[27]

Moral rights are established in the copyright laws of several countries, but are not universally supported and protected. The United States, for example, grants rights of attribution and integrity only to authors of works of visual art, and extends them to the end of the author's natural life.[37] However, German and French copyright law extend these moral rights to authors of all works and allow them to be transferred to heirs.[29] Moral rights allow the author to associate or disassociate herself from works, including derivative works, and, in the case of French law, to prevent release of or removal from public availability already published works. It is possible that moral rights will play a role in evaluating spoken word collections, particularly in the cases of unscripted or extemporaneous speech in oral histories, interviews, meetings, and the like, where it is perhaps more likely that a subject will wish to retract or withdraw.

4.4 Making the argument to digitise

Although several countries mandate or encourage legal deposit, it is not true that physical ownership confers ownership of the underlying intellectual content, unless a deed of gift or some other condition of acquisition explicitly transfers copyright along with the physical artifact. Therefore, even though national and depository libraries and archives have wonderful, unique and precious audio collections at their disposal,

they must look carefully at exemptions in the copyright law to make an argument in favor of digitizing, for most audio content is likely to be subject, in some degree, to copyright protections.

The "fair use" or "fair dealings" clauses are a natural starting point. Most countries have made some provision for reproducing copyrighted works for certain purposes. Those specifically mentioned include teaching, criticism, news reporting, and parody. In all cases, the language of the copyright law is non-specific as to the particulars. The United States copyright law's fair use clause cites "amount and substantiality of the portion used in relation to the copyrighted work as a whole" as a factor, but states that it must be balanced along with three other factors and offers no specifics about what a "substantial portion" might be.[38] In practice, fair dealings clauses are problematic. Their vagueness has led to self-censorship in many domains, including education, entertainment, and publishing. The content community has been successful in characterizing fair use as an archaic loophole.[22]

Specifics of copyright law vary from country to country, even among Berne convention signatories. It may be that a productive collaborative activity will be to establish some reasonable "acceptable risk" policies and practices, which need not be overly concerned with a narrow reading of any one copyright statute. As an inspiring example, the Australian National Archives recently decided to digitize archival materials and make them freely available, regardless of copyright status, to help overcome the "tyranny of distance":

> The approach adopted by the Archives was to look realistically at the nature of the material in question, and to look at the overriding purpose for which the Archives was planning to digitise and publish this material online. Generally the material in which copyright is held privately is of no commercial value. The records are mostly between 30 and 150 years old and, because of the passage of time, current copyright holders often cannot be identified or traced. The principal objective of our digitisation initiative is to fulfill our statutory function of encouraging and facilitating the use of archival material, and to that extent we have determined that it is in keeping with the spirit of the "fair use" provisions of our Copyright Act. On these grounds, the Archives felt that the public interest lay overwhelmingly in favour of proceeding with the initiative.[20]

Archives will set local policy based on the laws governing their country and the legal preferences of the parent institution, if any. Here is a standard list of questions to ask when considering whether or not a risk can be managed:[31]
--How old is the material?
--Was it produced for commercial or non-commercial purposes?
--Do deeds of gift or consent forms transfer any rights?
--Can a copyright holder be identified with a reasonable amount of effort?
--Can access be brokered in any way to mitigate potential concerns about worldwide distribution (only "thumbnail" equivalents or other deprecated versions delivered to the public at large, full network access granted to researchers who request certain materials, etc.)?
--Can digital liability insurance be obtained?
--Are there "safe harbor" provisions that require cease and desist notice before infringement action may be brought?

The conclusion regarding legal considerations is positive: clear violations of law are relatively easy to avoid. With due care, policies can and should be crafted to lower or eliminate legal barriers in the construction and use of online spoken-word collections, especially for activities which are non-commercial, educational and in the public interest. Archives should cooperate and define such due care through formulation of practical risk management policies and procedures.

5.0 COLLECTING, ARCHIVING AND PRESERVING CONTENT

5.1 Initial acquisition of content

In this report, preservation and archiving are discussed in reference to an organization that wishes to obtain and maintain content for the long term, and also wishes to make that content available to its community of users during the same period. We write from a particular perspective representing public archives, e.g., research libraries or national collections, and to some degree our ideas reflect organizations with a broad public responsibility. The technical concepts, however, apply as well to corporate, private, or for-profit archives.

One aspect of preservation and archiving concerns the initial acquisition of content. For example, persons with an interest in spoken-word collections are aware that extensive bodies of tape-recorded testimony resulting from oral history projects languish in small local libraries and historical societies. Similarly, many scholars who study language and dialect possess personal collections of sound recordings that have resulted from their research. There is clearly a public good to be served by placing these pre-existing, analog-format materials (or copies) in larger and more robust institutional archives.

Another corpus of interest to spoken language researchers is represented by born-digital content on the World Wide Web and other online contexts. This content is often ephemeral and short-lived. Archivists sometimes refer to this online content as *intangible* to distinguish it from digital content distributed in fixed media like compact disks. In recent years, the Library of Congress (US) and other national libraries have begun to collect and archive Web content, although to date this has generally not included sound recordings like radio webcasts. Those with an interest in spoken language, of course, will encourage the expansion of current collecting in order to secure this important cultural record for future generations. Producing organizations, e.g., broadcasters, share with public institutions the social responsibility of safeguarding this content.

In Europe, legal deposit legislation obliges publishers to place copies of printed matter in national libraries. Recent cases in France, Sweden, and Denmark have extended the definition of 'publication' to websites. This legislation is somewhat in advance of comprehensive Web archiving and preservation technology, but the action has launched a process of archiving Web content in Europe, including initial attempts to take audiovisual content from websites. National broadcast organizations like the BBC in Great Britain, and other major producers of media websites, are also actively involved in archiving content, including audio and video. Finally, the Internet Archive <http://www.archive.org/>, an independent non-profit organization in the United States, is attempting to archiving as much of the World Wide Web as is practical, and in a project shared with the Library of Congress has already made an impressive collection of broadcast coverage (audio and audiovisual) of the terrorist attacks in the United States on September 11, 2001.

The social and historical implications of existing analog recorded sound collections are striking. National archives and public broadcast archives in Europe have millions of hours of holdings, much of which features spoken language. Some sense of the extent can be seen in the results of a recent survey that identified on the order of ten million hours of sound recordings in Europe <http://presto.joanneum.ac.at/projects.asp#d2>. The bulk of this content -- estimated to be 80 percent -- is in analog form. The important fact about all analog material on tape is that it will perish within a few decades and that it is expensive to digitize (roughly US $100 per hour for preservation at archive standards). An order-of-magnitude estimate for all significant world holdings of audio material on analog formats is 100 million hours, with a preservation cost of US $10 billion.

What are the technological implications of the need to acquire spoken language content? Regarding the acquisition of physical materials like analog tape recordings, little needs to be said in the context of this report, devoted as it is to digital processing. These analog materials, once acquired, await the types of digital processing discussed in section 5.2 and following. Regarding the acquisition of intangible born-digital content, at least two technologies of interest to the field of spoken language research may be mentioned:

(1) Technologies to identify or filter content. For example, there are thousands of radio broadcasts on the Web, some of which are of high interest in terms of spoken language or in terms of content of broad public interest for the long term. Any archive planning an extensive Web harvesting activity will benefit from a system that finds and collects material selected according to the archive's guidelines. Today's software marketplace does not provide much in the way of sophisticated tools to accomplish this goal.

(2) Technologies to capture the transmitted bitstream. For example, many webcasts use one or another form of proprietary streaming media that require special technology at the point of capture to produce files that contain the stream. Today's marketplace provides only a few tools to accomplish this goal.

5.2 Processing, reformatting, or shaping content after acquisition. We examine in detail: shaping audio bitstreams, metadata, and special features of synchronization.

5.2.1 *Processing, reformatting, or shaping audio bitstreams.*
The core content element for those with an interest in spoken language is the sound recording itself. In the digital realm, this is represented by a bitstream, typically contained in a computer file. Spoken language processing is able to take advantage of a range of file and bitstream types, even when the quality as judged by an audiophile is only good. Inputs for such processing range from PCM (pulse code modulated) bitstreams in WAVE files to MP3 (MPEG 1, audio layer 3) bitstreams to RealAudio streaming audio, a currently widespread proprietary audio format.

Archivists with an eye on the long term, however, must be more discriminating than spoken language processing specialists. Archivists ask, "Which formats will endure and remain playable as time passes?" The Appendix to this report provides background on recent digital library discussions of preservation, including summaries of three approaches that have been proposed as preservation strategies: format migration, system emulation, and content normalization. The archivist must assess the bitstream or file in consideration of these strategies. Is the file migratable, i.e., can the fundamental information be moved into a new format in the future? Is the

file in a format for which we can expect playback-system emulations in the future? Does the archive have a system for normalizing digital content into a form that the archive proposes to maintain for the long term, and can this element be normalized into an appropriate form?

When reformatting older analog formats, most archives today select what they believe to be a migratable format for the master or "preservation" files they produce. Reformatting refers to the processes that are applied when content is "moved" from one media to another, e.g., the microfilming of books and newspapers. For audio, the preferred target format is a PCM sampling structure contained in a WAVE file format, e.g., WAVE and Broadcast WAVE (EBU Tech 3293: EBU Core Metadata Set for Radio Archives<http://www.ebu.ch/tech_t3293.html>). For practical reasons, derivative service or "access" files may also be produced, typically MP3 or some form of streaming audio, e.g., RealAudio.

Regarding born-digital files acquired by an archive, the format question is more challenging. For example, if a RealAudio stream is captured from a webcast, can the capturing archive count on the continued existence of playback software and/or emulations for the long term, or should this bitstream be reformatted into a different structure--e.g., as a PCM rendering--in hopes of increasing the likelihood of long-term playability? Will this kind of digital reformatting produce audio artifacts that mar the listen-ability of the recording? Is there a normalization strategy that may be helpful? Questions like these animate many digital library community discussions at this time; the spoken language community can contribute to this broader investigation by means of applications research or demonstration projects devoted to its particular type of content.

5.2.2. *Processing, reformatting, or shaping content: metadata.*
Members of different communities have various uses for the term 'metadata'. In our role as community representatives, we offer the following usage overview. Metadata includes:

A. High level "bibliographic" or "tombstone" information. This is the type of information one would expect to see in, say, a library catalog: identifier, title, creator, subject, date of recording, abstract of the content, identification of the responsible archive, etc. Many librarians refer to this as descriptive information. Compare to D below.

B. Information about the structure and organization of a multipart digital object. This is the type of information found in the METS (Metadata Encoding and Transmission Standard) Structural Map. For example, the Structural Map identifies the segments in an audio stream or series of files, or the individual page images in the reproduction of a book. Other formats also encode this type of metadata, e.g., MPEG-7 <http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm> includes the ability to document program segments and their relationships, while MPEG-21 <http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm> includes a content declaration, which is something like a "shipping manifest."

C. Administrative information of a variety of types. A good example of the cross section of administrative metadata may be seen in the METS standard: (1) technical metadata about the bitstream files in the object at hand, e.g., sampling frequency and word length; (2) source metadata, i.e., information similar to (1) but about the analog or digital elements that were reformatted into the current object; (3) digital

provenance or process metadata, i.e., information about how the bitstream files were created or reformatted; (4) rights or restrictions metadata, information that can be used to support access management systems; and (5) what is called behavior metadata, information that pertains to how to "play" the content at hand, comparable in some ways to "methods" in other computer applications.

D. Transcriptions of spoken language recordings. We focus on instances in which transcripts--typically produced by automated systems--are supporting and supplementary to the recording proper, and thus may be seen as metadata rather than as data. The authors recognize, however, that in some contexts transcripts may be defined as alternate manifestations of the work itself, i.e., an element co-equal with a sound recording, or that they may be "works of their own," a phenomenon strikingly represented by the Congressional Record, the official record of the debates of the U.S. Congress. The Record is an edited and amended version of the words actually spoken on the floor of the U.S. House of Representative and U.S. Senate.

5.2.3. *Special features of synchronization*

At one level, there are no mysteries regarding synchronization of audio and transcript. Sound recordings are time based and their digital-audio representations, broken into the exceedingly fine (a few microseconds) divisions of samples or frames--can be dynamically aligned with transcripts that are marked in some way that indicates the elapsed time of a given phone, word, or phrase. One source of complexity, however, results from the need to synchronize transcript-based lattices or trellises (multiple-hypothesis structures) with the audio stream. A second source of complexity arises from the practices used in some research or access environments. For example, time may be non-linearly "compressed" in an audio stream to facilitate the efficient delivery of an extensive chunk of sound. If the transcript is synchronized to the original, uncompressed recording, how shall we re-synch the compressed version to the transcript?

For the archivist, the various issues in synchronization lead to a desire for conventions or standards that permit the exchange of digital objects between archives. These standards, to be sure, will also facilitate data exchange between individual researchers. Standardization will also lead to the creation of shared tool sets, thereby reducing the need to customize software.

5.3 Packaging content for archiving and access

5.3.1 *Content packages and the preservation repository*

The Appendix outlines current thinking in the digital library community regarding repositories to archive content for the long term. A critical element is the Open Archival Information System (OAIS) reference model, whose structure expresses phases of the digital content life cycle. A key feature of the OAIS model is the content or information package, conceived of as an object that bundles data and metadata for the sake of content management. Content packages include:

--Files or bitstreams that represent the content as content, e.g., a WAVE file that reproduces sound.

--Metadata, which may be further subdivided as indicated above.

--Encapsulation schemes, e.g., the use of UNIX "tar" files to bind a package as it is moved within or between systems.

5.3.2. *Standards and best practices in formatting and packaging*

Archivists and librarians are developing a number of standards and guidelines pertaining to digital content. For sound archivists, for example, high level ethical, strategic, and practical guidance is provided by the International Association of Sound and Audiovisual Archives (IASA) in their document, "The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy" (IASA-TC 03 Version 1, February 1997). Another example is the OAIS reference model itself, now an ISO standard. Most relevant to the purposes of this report are the elements described in the following sub-sections. The unresolved aspects of these elements do not require laboratory research but rather the establishment of conventions to aid in the preparation and structuring of content. These conventions and associated practices are essential to the practice of archiving and thus to the long-term availability of spoken language content.

5.3.2.1 *Audio files.* Assuming that the widely used PCM sampling structure is suitable for spoken language master files, are certain sampling rates or bit depths (word lengths) preferred for spoken language recordings?

Should the spoken language community make a recommendation of MP3, RealAudio, or other formats for the dissemination of content ("service" files)?

Are there special features or structures pertaining to spoken language recordings that are omitted from existing standards, e.g., for WAVE, Broadcast WAVE (BWF), or MP3 file formats?

As the spoken language community develops guidelines, it should consider existing practices and standards adopted by various technical bodies. For example:
-- EBU Recommendation Rdra-2001: Digitisation of programme material in radio archives
-- EBU Technical Recommendation R84-1996: Word length, sampling rates and auxiliary information in digital systems used for high-quality audio production
-- IEC 60268: Sound system equipment

5.3.2.2 *Metadata other than transcripts.* Metadata may be embedded within files or held in associated databases, XML documents, and the like. These categories can be expected to overlap. Consideration that guide practitioners to select one or another location for various types of metadata include the following:
-- How often will metadata change (e.g., information about the use or manipulation of the audio as in a broadcast production setting)?
-- Is there need for fast searching of metadata, leading to use of an external database, and to the indexing of the metadata?
-- Are there security considerations?

In order to link the audio with external metadata, it is essential to have an unambiguous identifier as one of the embedded metadata elements. There are many schemes for generating identifying numbers. In the broadcast world, the BWAV format uses the USID (EBU Technical Recommendation R99-1999; 'Unique' Source Identifier (USID) for use in the OriginatorReference field of the Broadcast Wave Format), and in the cinema and broadcast video world (which includes sound), there is some support for the SMPTE UMID (SMPTE STANDARD 330 for Television - Unique Material Identifier (UMID; The Society of Motion Picture and Television Engineers. - Copyright 1999; Universal identifier as link between embedded and external metadata)). Other organizations, like the Library of Congress, have begun

to use the handle form of a URI as a persistent identifier for its digital content <http://www.handle.net/introduction.html>.

If an audio file is part of a coherent digital-content object with other elements, which parts of any file-embedded metadata ought to be "brought to the surface" in the object?

How shall we identify languages and/or the markup and encoding of language representations in text? References to standards on language coding include:
- - Internet Engineering Task Force - Tags for the Identification of Languages RFC 1766. <http://www.ietf.org/rfc/rfc1766.txt>
- - ISO 639 - Codes for the representation of names of languages (Registration Authority). <http://lcweb.loc.gov/standards/iso639-2/>

5.3.2.3 *Transcripts as metadata.* Are there accepted practices in rendering transcripts that should be recommended to archives and incorporated into demonstration projects? For example, does the spoken language community find that the MPEG-7 Spoken Content Description Scheme (DS) is worthy of testing or use? Are there alternatives? The MPEG-7 documentation states, "The Spoken Content DSs are a representation of the output of Automatic Speech Recognition (ASR). The SpokenContentLattice represents the actual decoding produced by an ASR engine, whereas the SpokenContentHeader contains information about the recogniser itself and the people (or "Speakers") being recognised."

Regarding the representation of synchronization, what alternatives may be considered? Synchronization is implicit in the MPEG-7 Spoken Content Description Scheme noted above. Another option is the W3C (World Wide Web Consortium) Synchronized Multimedia Integration Language (SMIL, pronounced "smile").

How shall we identify languages and/or the markup and encoding of language representations within a text, including the vexing matter of such things as sentences that include a single "foreign" word?

Humanities scholars have invested much effort in developing the Text Encoding Initiative (TEI) markup language and associated conventions in order to exchange textual renderings of printed and written documents. Considering that many of the spoken language recordings with enduring interest to society, e.g., oral histories, are important documents for the humanities, is there merit to investigating an expansion of the TEI schemes to spoken language transcripts, including a recommended approach for indicating elapsed time?

Are there actions or conventions that will make transcripts usable by researchers from multiple disciplines? For example, will renderings that feature the "annotation graphs" employed by workers in the spoken language processing and speech recognition communities be comprehensible or helpful to members of other communities, e.g., oral historians, folklorists, and cultural anthropologists? Or what might specialists in the latter fields do to make their content more useful to the spoken language processing and speech recognition communities?

5.3.2.4 *Packaging standards.* What should spoken-language-content archivists recommend regarding the selection of standards that perform packaging functions, e.g., METS, MPEG-21 and/or MPEG-7?

To what degree can file-specific metadata serve to package content? There is minimal metadata in WAV files, and a bit more in BWAV. There are proposals to add an XML chunk to BWAV, allowing for an indefinite extension of metadata embedding. It is by no means clear that this is a principled approach, because the clear alternative is to segregate WAV data from XML data, and link them with an identifier. None of these approaches fully address the problem of the overall package or wrapper.

Packaging in the multimedia production environment has recently emerged, although this is oriented to data exchange more than to long-term data preservation. There is a need in production to handle complex multimedia structures, including file storage, electronic exchange, and broadcasting of the final product. A certain level of consensus has been reached within the broadcast, cinema and IT industries that favors the use of the AAF (Advanced Authoring Format <http://www.aafassociation.org/>). A somewhat simplified version of AAF has been developed and is being standardized via MPEG and SMPTE, the MXF (Media Exchange Format <http://www.g-fors.com>). MXF certainly performs packaging functions, but it supports much else as well, and may be overkill as a recommendation for an audio+metadata package.

5.3.2.5 *Other.* Finally, we ask if there are there special classes of metadata pertaining to content of interest to the spoken language community not addressed by any other standards? Is there a process that will describe these classes and take the actions that may be needed to establish a standard? To what degree will practices in this area address the concerns expressed in Steven Bird's and Gary Simons's article "Seven Dimensions of Portability for Language Documentation and Description" <http://arxiv.org/abs/cs.CL/0204020>.

5.4 Sustainability

Digital content in technical terms is sustainable. Sustainability in financial terms, however, is another matter. It is a focus of concern, although not specifically within the group's expertise. What business case can be made to support the existence of a spoken language archive? We note the following aspects to this topic:

-- The acquisition and archiving of content of evident social value should be supported by society, i.e., government libraries and archives that are funded by taxes. This includes content of interest to scholarship, like oral histories, selected lectures by academics, judicial proceedings, and the records of government, e.g., recordings of the deliberations of political bodies.

-- Content owners will sponsor content of commercial interest. For example, many broadcast recordings have continuing value in commerce, e.g., as material for rebroadcast or for sale to others engaged in program production. This content is likely to be preserved, although society should encourage this preservation and stand by to receive material when commercial interests retire it.

-- Content of interest to the law enforcement and national security communities will be archived by its members and in some cases may pass into public archives in the future. Some of this content has evident social value. The business case for long-term preservation should be considered.

-- The generation of content intended for specific spoken language research purposes is often funded for the purposes of that research. Some of this content has

evident social value, and the business case for long-term preservation should be considered.

ACKNOWLEDGMENTS: This report is an amalgam resulting from ideas identified, discussed and debated by the working group members over the last eight months. We would like to take this opportunity to thank NSF and DELOS for their financial support enabling the working group members to meet, identify, disagree and debate the topics central to this report. Finally, we express our gratitude to the staffs at the University of Sheffield, TNO Delft, and Northwestern University for coordinating complicated meeting and planning schedules with grace and efficiency.

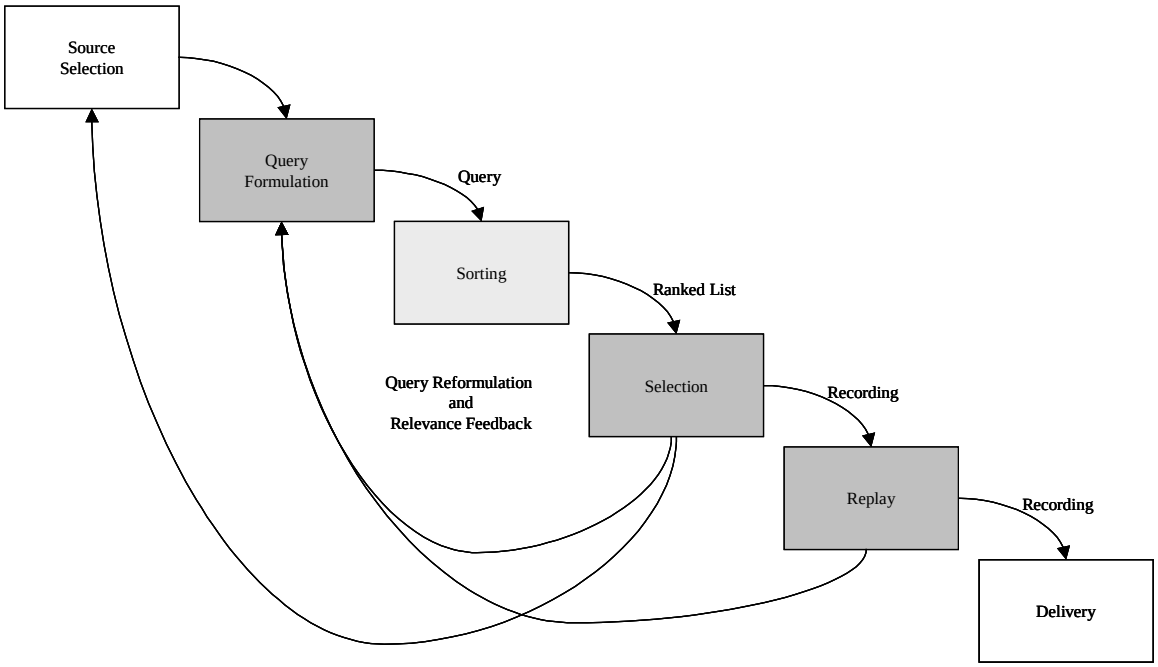

Figure 1: Interactive Search Process

REFERENCES

[1] Berne Convention, Article 2bis

[2] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani. Cross-task portability of a broadcast news speech recognition system. Speech Communication, 38(3-4):335-347, 2002.

[3] S. S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, 1998.

[4] Coretex. EU project. EU Project See coretex.itc.it, 2000-2003.

[5] Council of European Communities. Council Directive 93/98/EEC of 29 October 1993 Harmonizing the Term of Protection of Copyright and Certain Related Rights, 29 October 1993. Accessed on March 14, 2003: <http://europa.eu.int/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=en&numdoc=31993L0098&model=guichett>

[6] N. Deshmukh, A. Ganapathiraju, R. J. Duncan, and J. Picone. Human speech recognition performance on the 1995 csr hub-3 corpus. In Proceedings ARPA Speech Recognition Workshop, pages 129-134, Harriman, NY, February 1996.

[7] G. R. Doddington. Speaker recognition - identifying people by their voices. Proceedings of the IEEE, 73(11):1651-1664, 1985.

[8] W. J. Ebel and J. Picone. Human speech recognition performance on the 1994 csr spoke 10 corpus. In Proceedings ARPA Spoken Language Systems Technology Workshop, pages 53-59, Austin, TX, January 1995.

[9] Employee Privacy: Computer-Use Monitoring Practices of Selected Companies. Washington, D.C.: United States General Accounting Office, September 2002. Accessed on March 13, 2003: <http://www.gao.gov/new.items/d02717.pdf>; Privacy and Human Rights 2002, pages 90-91.

[10] European Parliament. Directive 2001/20/EC of the European Parliament and of the Council. 4 April 2001. Accessed on March 13, 2003: <http://europa.eu.int/eur-lex/pri/en/oj/dat/2001/l_121/l_12120010501en00340044.pdf>

[11] European Parliament. Directive 2001/29/EC of the European Parliament and of the Council. Accessed on March 14, 2002: <http://www.patent.gov.uk/copy/notices/pdf/implement.pdf>

[12] The Federal Judiciary (United States). Wiretap Reports. Accessed on March 13, 2003: < http://www.uscourts.gov/wiretap.html>

[13] Garofolo, J.S., Auzanne, G.P., Voorhees, E.M., "The TREC Spoken Document Retrieval Track: A Success Story," Proceedings of the Recherche d'Informations Assiste par Ordinateur: ContentBased Multimedia Information Access Conference , April 12-14, 2000. <http://citeseer.nj.nec.com/garofolo00trec.html>

[14] H. Gish and H. Schmidt. Text-independent speaker identification. IEEE Signal Processing Magazine, October:18-32, 1994.

[15] R. Glover and A. Worlton. Trans-national employers must harmonize conflicting privacy rules. In The Metropolitan Corporate Counsel, Mid-Atlantic Edition. page 20 (November 2002)

[16] Information Commissioner (United Kingdom). CCTV code of practice. July 2000. Accessed on March 13, 2002, available at: <http://www.dataprotection.gov.uk/ >

[17] D. S. Karjala. Chart showing changes made and the degree of harmonization achieved and disharmonization exacerbated by the Sonny Bono Copyright Term extension Act (CTEA). May 15, 2002. Accessed on March 14, 2003: <http://www.law.asu.edu/HomePages/Karjala/OpposingCopyrightExtension/legmats/HarmonizationChartDSK.html>

[18] L. Lamel, J.-L. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language, 16(1):115-229, 2002.

[19] L. Lamel, F. Lefevre, J.-L. Gauvain, and G. Adda. Portability issues for speech recognition technologies. In Proceedings of HLT 2001, pages 9-16, San Diego, CA, 2001.

[20] T. Ling. Why the archives introduced digitisation on demand. RLG Diginews 6(4)(August 15, 2002). Accessed on March 14, 2003: <http://www.rlg.org/preserv/diginews/diginews6-4.html#feature1>

[21] R. P. Lippmann. Speech recognition by machines and humans. Speech Communication, 22(1):1-15, 1997.

[22] J. Litman. Digital copyright, page 84. Prometheus Books, Amherst, NY, UDA, 2001.

[23] J. M. Naik. Speaker verification: A tutorial. IEEE Communication Magazine, January:18-32, 1994.

[24] Patent Office (United Kingdom). Copyright History. Accessed on March 13, 2003: <http://www.patent.gov.uk/copy/history/>

[25] Privacy and Human Rights 2002. Electronic Privacy Information Center (EPIC) and Privacy International, 56-57, Washington, D.C., 2002.

[26] L. R. Rabiner and B. Juang. An introduction to hidden markov models. IEEE Acoustics Speech and Signal Processing ASSP Magazine, ASSP-3(1):4-16, 1986.

[27] M. T. Sundara Rajan. Moral rights and copyright harmonisation: prospects for an "international moral Rright"? In 17th BILETA Annual Conference. April 2002. Accessed on March 14, 2003: <http://www.bileta.ac.uk/02papers/sundarajan.html>

[28] République Française. Law on Author's Rights and on the Rights of Performers, Producers of Photograms and Videograms and Audiovisual Communication

Enterprises[translated title]. Accessed on March 14, 2003 and available in English translation from: <http://clea.wipo.int/>

[29] République Française. Law on Author's Rights [translated title], Copyright (Part I), Code (Consolidation), 01/07/1992 (03/01/1995), No. 92-597 (No. 95-4), Title II, Art. L. 121-1.

[30] D. Reynolds. Speaker identification and verification using gaussian mixture speaker models. Speech Communication, 17:91-108, 1995.

[31] Risk Management Suggestions. Multimedia & Web Strategist 5(4)(January 1999). Accessed March 14, 2003, in Lexis-Nexis Academic Universe.

[32] A. E. Rosenberg. Automatic speaker verification: A review. Proceedings of the IEEE, 64(4):475-487, 1976.

[33] A. E. Rosenberg and F. K. Soong. Recent research in automatic speaker recognition. In Advances in Speech Signal Processing, chapter 22. M. Dekker, Inc., New York, NY, 1992.

[34] L. E. Rothenberg. Re-thinking privacy: peeping toms, video voyeurs, and failure of the criminal law to recognize a reasonable expectation of privacy in the public space. American University Law Review, 49:1127, June 2000.

[35] G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461-464, 1978.

[36] The search & seizure of electronic information: The law before and after the USA PATRIOT act. Washington, D.C.: Wiley, Rein and Fielding, LLP. Accessed on March 13, 2003: <http://www.ala.org/washoff/matrix.pdf>

[37] Title 17 U.S. Code, Chapter 1, Section 106A

[38] Title 17 U.S. Code, Chapter 1, Section 107

[39] A. Tritschler and R. Gopinath. Improved speaker segmentation and segment clustering using Bayesian Information Criterion. In Proceedings of the 6th European Conference on Speech Communication and Technology, pages 679-682, Budapest, Hungary, 1999.

[40] D. A. van Leeuwen, L. G. van den Berg, and H. J. M. Steeneken. Human benchmarks for speaker independent large vocabulary recognition performance. In Proceedings of the 4th European Conference on Speech Communication and Technology, pages 1461-1464, Madrid, Spain, 1995.

[41] S. Young and G. Bloothooft, editors. Corpus Based Methods in Language and Speech Processing. Kluwer, Berlin, Germany, 2000.

[42] U.K. JISC Data Protection Principles. Accessed on March 13, 2003: <http://www.jisc.ac.uk/legal/index.cfm?name=lis_dp_prin> and U.S. NIH Office of Human Subjects Research. Accessed on March 14, 2003: <http://206.102.88.10/ohsrsite/>

43 World Intellectual Property Organization (WIPO). Berne Convention for the Protection of Literary and Artistic Works. Accessed on March 14, 2003: <http://www.wipo.int/treaties/ip/berne/index.html>

ILLUSTRATIVE AND EXEMPLARY RESOURCES
European Language Resources Association (ELRA)
<http://www.icp.inpg.fr/ELRA/home.html>
Linguistic Data Consortium (LDC)
<http://www.ldc.upenn.edu/>
Evaluation:
NIST Spoken Natural Language Processing <http://www.itl.nist.gov/div894/894.01>
Projects:
Coretex Improving Core Speech Recognition Technology (EC project)
<http://coretex.itc.it>
ECHO European Chronicles Online (EC project)
<http://pc-erato2.iei.pi.cnr.it/echo/>

Appendix. Content Preservation from a Digital Library Perspective

App.1 Preservation problems and strategies. The problem of preserving content in digital form has received widespread attention in the library and archive community in recent years, starting with dramatic statements of the problem in the early 1990s, moving to solution proposals in the early 2000s. The solution statements call for of actions on multiple fronts--technical, policy and political, organizational--and some of these actions are beginning to come into view today.

The first problem statements employed familiar examples, e.g., old 5.25-inch diskettes that cannot be played in the new 3.5-inch drives in new computers, with added remarks about reading WordStar documents in a Microsoft Word environment. One of the most widely read and carefully developed expressions of these ideas was Jeff Rothenberg's article "Ensuring the Longevity of Digital Documents" *(Scientific American*, January 1995).

App.1.1 Format migration. In the Library community, the solution statements began by introducing a pair of terms that were at first seen as both distinct and opposite: *migration* and *system emulation*. The former was articulated in the 1996 report by Don Waters and John Garrett, *Preserving Digital Information: Final Report and Recommendations* <http://www.rlg.org/ArchTF/index.html>. In that report and in subsequent discussions, two meanings for *migration* have emerged. The first, which many call *refreshment*, refers to the movement of an unchanged bitstream from one medium or device to another, more or less what happens when you move your files unchanged from your old computer to its new replacement. This is essential to the proper workings of any computer system--especially big ones--but less fraught with issues than the other type of migration. The second meaning, which some call *format migration*, refers to the movement of content or data from one bitstream structure (approximately "file format") to another, more or less what happens when you take your WordStar document and convert it to Word. As the example suggests, this type of migration is usually contemplated because of the obsolescence of the existing bitstream structure or, more accurately, the obsolescence of the software tools needed to render that bitstream into a form comprehensible by humans.

Formats established by such standards groups as ISO (International Organization for Standardization), NISO (National Information Standards Organization), W3C (World Wide Web Consortium), and I3A (International Imaging Industry Association) are very likely to be migratable. Some widely deployed, openly documented industry formats industry are also likely to be migratable, e.g., the TIFF (Tagged Image File Format) format, developed by Aldus, Inc., now Adobe, Inc. TIFF is an example of an "open format," marked by minimal proprietary information. There exists what might be called an "openness spectrum," with degrees of protection applied by industry developers. For example, the PostScript page description language (also from Adobe, Inc.) could be described as "proprietary and open." Very comprehensive and fully public documentation is available for some proprietary content formats, e.g., TIFF.

Other factors also increase the feasibility of migration. For example, machine-readable texts will be easier to migrate when the character set is known, especially when the character sets follow ISO/NISO standards, with the UNICODE international character set seemingly the best choice for long term content preservation. But

experts in language correctly warn that the encoding of important features of language requires more than just a character set.

Migratability is increased when thorough documentation for a given format is public and widely available, and will be further increased when the facts of the creator's use of the format (e.g., facts about the customization of the creating software application) have also been documented. Although never a sufficient condition, migratability may be more plausible when a format is widely deployed and implemented, or when it can be opened in software applications from a broad range of companies. Migration will succeed when the significant or essential features of content are maintained or when a given unit or subunit of that content is migrated from an obsolescent format to its replacement.

The prospect for migrating audio files varies according to the format. The core "sound" content in many audio files consists of *pulse code modulated* (PCM) data in a linear stream. Linear PCM data, irrespective of sampling rate, word length, method of packing data into bytes and left-to-right or right-to-left arrangement of bits and bytes, can be decoded by relatively simple trial-and-error, and we can expect this to be the case indefinitely. PCM is in this sense a 'natural' representation for audio, and has very good long-term prospects regardless of the remaining problems of format migration. The situation is very different for compressed audio, particularly proprietary compression, and the format migration problem could be severe for current common formats such as MP3 and Real Audio, and this problem may well develop in no more than another decade.

> If an audio file or an associated file includes extensive metadata, such as time-aligned annotations or transcripts or multiple-hypothesis 'lattice' transcripts, then the file or file group would indeed have a complex structure and would require a complex approach for 'format migration.' It is likely to be the metadata that would cause difficulty, not the actual bits representing the audio, providing the audio data is in simple PCM coding. By confining the metadata to a contiguous area within the file, or with little or no metadata, the remaining data will have a very high likelihood of surviving any format migration unscathed.

App.1.2 System emulation. The shortcomings of format migration contributed to the advocacy of system emulation by Jeff Rothenberg in his report *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* <http://www.clir.org/pubs/abstract/pub77.html>, and in similar recommendations by others. Rothenberg pointed out that format migrations change the look and feel of documents. Anyone who has migrated heavily formatted documents, e.g., word processing files containing such elements as tables or automatically numbered outlines, has experienced the damage inflicted by conversion (i.e., migration) programs and the consequent need to edit the new version to get back what was lost. This led Rothenberg to argue in favor of leaving the document unchanged and finding ways to emulate the systems that rendered it, e.g., and emulation of WordStar running on an emulation of Windows 3.1 running on an emulation of an Intel 286 chip.

Such true emulations may exist only for a few system types and may be costly to maintain. In some cases, organizations will develop "pretty good" emulators that render content with minimal changes in appearance or functionality. The concept of what might be called *levels of preservation*--which apply equally to migration and

emulation--is the topic of Paul Wheatley's very helpful paper titled "Migration - a CAMiLEON discussion paper" (<www.ariadne.ac.uk/issue29/camileon>; also <www.personal.leeds.ac.uk/~issprw/camileon/migration.htm>; see Section 2 for Wheatley's discussion of various levels of preservation). In a recent study, the matter of maintaining look and feel has been questioned. An activity under the direction of Margaret Hedstrom at the University of Michigan took a computer game and compared user reactions to migrated and system-emulated versions; see "Emulation vs. Migration: Do Users Care?" by Margaret Hedstrom and Clifford Lampe <http://www.rlg.org/preserv/diginews/diginews5-6.html#feature1>. In both cases, there were changes in look and feel (in the words of the article, "change happens") and it was a toss-up as to which approach provided a more satisfactory outcome for fans of the game.

App.1.3 Content normalization. Several digital library commentators argue that the preservation of content requires a judicious mix of format migration and system emulation, sometimes referred to as *normalization.* A normalization scenario might work like this:
--A library acquires content in a less preservable format.
--The content is normalized (i.e., *migrated*) to a format that is well understood by that library.
--The library creates a hardware and/or software system that can render the normalized content.
--The library commits to re-creating the rendering system for the long term, i.e., committing to *emulate* the "original" system designed by the library to render the normalized content.

Variations on the normalization approach have been developed or discussed in the digital library community. One example is the *Persistent Archive* design being tested within what is called the Data-intensive Computing Thrust at the supercomputer center at the University of California, San Diego, also known as the National Partnership for Advanced Computational Infrastructure <www.npaci.edu/Thrusts/DI/index.html>. The persistent archive approach migrates content into what are called *persistent objects* for long-term management. One version of the system is being tested at the National Archives and Records Administration (documents available from <www.sdsc.edu/NARA/Publications.html>). A second example is outlined in the article "A Project on Preservation of Digital Data" by the IBM computer scientist Raymond Lorie (*RLG DigiNews*, June 15, 2001; **<**http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2>). Lorie proposes that digital library organizations create *Universal Virtual Computers* to their own specifications and then reformat content for that computer environment. The organizations are thereby well positioned to update the virtual computer over time as needed. Some comparisons may be made between this approach and ideas associated with the Java virtual computer promoted by Sun Systems and others. Third, the National Library of Medicine (NLM) preserves a number of medical e-journals by reformatting them from their native markup language (typically based in a particular XML DTD [Extensible Markup Language Document Type Definition] or schema) into an NLM-developed XML schema, thus simplifying management for the long term.

Emulation or normalization issues for the core "sound" content in linear PCM encoded digital audio files will be simpler than for documents or computer games or similar complex artifacts. This representation of the audio signal is inherently simple: a

sequence of samples. Thus the development and maintenance of a normalized format for audio is quite feasible. Indeed, linear PCM is already largely accepted as the 'normal' form for a digital audio signal. In addition, for over fifteen years the WAVE file format has been a de-facto standard for audio files, and this should remain the case for at least that long into the future.

As stated above regarding format migration, the matter become more complex when the audio stream has been compressed and is stored as, say, MP3 or RealAudio data. In addition, if complex metadata is associated with a digitally-represented audio signal - - for example, a marked-up transcription - - then all the foregoing discussion of emulation and/or normalization applies, with the added complexity that the document may have a structure for time-registration against the audio signal.

App.2 The Digital Repository and Content Packages. The preceding section outlines part of the digital library community's current thinking about keeping content alive over the long term. The focus on bitstreams or files, however, leaves unsaid the considerable matter of *how* content management might be accomplished, and by whom. One answer is packed into the term *digital repository*. Many writers, including the authors of the report *Attributes of a Trusted Digital Repository* (Research Libraries Group [RLG] and the Online Computer Library Center [OCLC], draft August 2001), use the term to name an organization in the fullest sense, i.e., both the staff and the technical systems employed to manage digital content. In the words of the report:

> A reliable digital repository is one whose mission is to provide long-term access to managed digital resources; that accepts responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users; that designs its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it; that establishes methodologies for system evaluation that meet community expectations of trustworthiness; that can be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly; and whose policies, practices, and performance can be audited and measured.

App.2.1 The OAIS reference model. Regarding the technical systems, many in the digital library community have embraced the Reference Model for Open Archival Information Systems (OAIS). The model is introduced on the National Air and Space Administration website in these words (1999):

> ISO has undertaken a new effort to develop standards in support of the long term preservation of digital information obtained from observations of the terrestrial and space environments. ISO has requested that the Consultative Committee for Space Data Systems Panel 2 coordinate the development of those standards. <http://ssdoo.gsfc.nasa.gov/nost/isoas/>. (The document proper is at <http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf>)

At this writing it would be fair to say that the digital library community's embrace of the OAIS reference model has been somewhat theoretical; there is much discussion and some activity to implement portions of it. Several features have been especially appealing to this community, two of which are evident in the following simplified diagram. One is the modular nature of the architecture, which is in turn suggestive of what might be called a "content life cycle." "Production" and "ingestion" are shown

at the left, indicating that content is first shaped by its makers and/or reshaped by the repository organization upon arrival, to make it fit for long term archiving. Content is submitted by producers as Submission Information Packages (SIPs) for ingestion; then reshaped by the repository into Archival Information Packages (AIPs) for storage and preservation. Finally, when delivery is called for, content is once again reshaped by the repository into Dissemination Information Packages (DIPs) for presentation to end-users.

App.2.2 Content packages. The second feature of interest--in keeping with this modular view--is the notion that digital content can be seen as a "package" (for library preservation specialists this reminds us of how a microfilm of a book keeps all the pages together) and that the package may differ when submitted to the repository systems, when managed by the repository for the long term, and when delivered to a customer. As this concept suggests, well-designed and standardized content packages also support the exchange of objects between repositories and their presentation to researchers.

How does this map to our notions of normalization, migration, and or system emulation? The discussions in the community often associate normalization with ingestion, i.e., when content is readied for archiving, although there may also be *ex post facto* normalizations as part of management of content in archival storage. Migration presumably will be an action that takes place as content is managed in archival storage. System emulations as a requirement will be part of archival storage management as well, and their availability and application will occur in the access module, when content is presented to users.

A third feature of the OAIS reference model that has been of interest to the library and archiving community concerns the four types of metadata that must be associated with content. Two metadata categories, packaging and descriptive information, are relatively self-explanatory. Packaging information is a kind of shipping manifest that binds the package together into a single identifiable unit. Descriptive information is the old librarian's friend, information that can be indexed to support resource discovery. Content information names the content-carrying bitstreams themselves and the information needed to render and/or interpret the package. Preservation description information carries the metadata that is needed to preserve the content information.

Clearly these latter categories are crucial and they have been analyzed by working groups sponsored by the Research Libraries Group and the Online Cataloging Library Center. (See the listing at <http://www.rlg.org/longterm/index.html>. The listing includes pointers to *Preservation Metadata for Digital Objects: A Review of the State of the Art* (PDF, January 2001; <http://www.oclc.org/research/pmwg/presmeta_wp.pdf>), *Part I: A Recommendation for Content Information* (PDF, October 2001; <http://www.oclc.org/research/pmwg/contentinformation.pdf>), and *Part II: A Recommendation for Preservation Description Information* (PDF, April 2002; <www.oclc.org/research/pmwg/pres_desc_info.pdf>).

The issues that bear on or about shaping packages and retaining appropriate types of *content information* and *preservation description information* in the realm of spoken language data are nicely brought to life in Steven Bird's and Gary Simons's article "Seven Dimensions of Portability for Language Documentation and Description" <http://arxiv.org/abs/cs.CL/0204020>. This article indicates the many

challenges needed to produce (on the one hand) the needed metadata and (on the other hand) the supporting systems for emulation, migration, or normalization of language data in an OAIS package.

App.3 Notes on standards. The preceding discussion highlights a number of zones in which standardization is welcome. In some of these, true standards or open industry standards have been established or are proposed. In others, standards development is needed. The following overview is broad; within this context will be found certain zones in which the spoken language community can make a special contribution.

For repositories, the report *Attributes of a Trusted Digital Repository* (Research Libraries Group [RLG] and the Online Computer Library Center [OCLC], draft August 2001, pp. 12-14), outlines the attributes of policies, practices, and performance that can be audited and measured:
--Administrative responsibility
--Organizational viability
--Financial sustainability
--Technology suitability
--System security
--Procedural accountability

The report discusses various models and approaches for certifying complex technical and administrative entities, e.g., the ISO 9000 process, noting that no such models or approaches have been established for digital-content repositories (pp. 14-17). The report also calls attention to the unanswered need to establish auspices and authoritative organizations for repository certification. The Working Group believes that repository system design and certification are matters beyond the scope of a spoken language research effort.