Chapter 11

Managing Multimodal Data, Metadata and Annotations: Challenges and Solutions

Andrei Popescu-Belis
Idiap Research Institute
Martigny
Switzerland

11.1 Introduction

The application of statistical learning to multimodal signal processing requires a significant amount of data for development, training and test. The availability of data often conditions the possibility of an investigation, and can influence its objectives and application domains. The development of innovative multimodal signal processing methods depends not only on raw data recorded from various sources, but more specifically on high added-value data that is accompanied by ground-truth metadata or annotations. This information, generally added or validated manually, is used for supervised learning and for testing.

The goal of this chapter is to outline the main stages in multimodal data management, starting with the capture of multimodal raw data in instrumented spaces (Section 11.3). The challenges of data annotation – mono or multimodal – are discussed in Section 11.4, while the issues of data formatting, storage and distribution are analyzed in Section 11.5. In particular, Section 11.5.3 provides a discussion of the access to multimodal data sets using interactive tools, and more specifically of some meeting browsers that were recently developed to access multimodal recordings of humans in interaction.

The chapter starts, however, by introducing in the following section (11.2) an important conceptual distinction between metadata and annotations, and then surveying a number of important projects that have created large multimodal collections, and from which best practice examples will be drawn throughout the chapter.

11.2 Setting the stage: concepts and projects

Multimodal signal processing generally applies to data involving human communication modalities in two different types of settings: human-human vs. human-computer interaction (HCI). In the first setting, the goal of multimodal processing is to grasp and abstract certain aspects of human interaction. Processing can occur either in real time, or offline, after the interaction took place, for instance in order to facilitate search in multimodal recordings, or to draw conclusions regarding human performance in recorded sessions. In the second setting (HCI), the goal of multimodal processing is mainly to react in real time to a human's multimodal input to a computer, by generating an appropriate reaction (behavior) from the computer. Although multimodal data is certainly not absent from research and development in the HCI setting (e.g. as recordings of Wizard-of-Oz experiments aimed at modeling users' behavior), it is mainly in the first area that large quantities of annotated multimodal data are especially useful, for training and testing machine learning tools.

In this chapter, we use the term *corpus* to refer to a meaningful set or collection of data. A multimodal corpus is a set of data files (raw data, metadata and annotations) containing recordings of humans in interaction according to several modalities. The term *corpus* has been initially used to define a set of texts put together for a meaningful purpose, and therefore exhibiting a certain form of coherence (of topic, of style, of date, etc.). While the fields of speech and language processing commonly use corpora in their data-driven investigations, other fields simply refer to *data sets* or *collections*. Several recent initiatives (see Section 11.2.2) use however the word *corpus* to refer to their multimodal data sets, and corpus distributors such as the Linguistic Data Consortium have on offer a significant set of multimodal corpora. A series of workshops on this topic exists since 2000 (see http://www.multimodal-corpora.org).

11.2.1 Metadata vs. annotations

Throughout this chapter, we make reference to an important distinction: we define *annotations* as the time-dependent information which is abstracted from input signals, and which includes low-level mono or multi-modal features, as well as higher-level phenomena, abstracted or not from the low-level features. For instance, speech transcription, segmentation into utterances or topical episodes, coding of gestures, visual focus of attention, or dominance will be called here annotations.

Conversely, we define *metadata* as the static information about an entire unit of data capture (e.g. a session or a meeting), which is not involved in a time dependent relation to its content, i.e. which is generally constant for the entire unit. For instance, examples of metadata items are: date, start and end time, location, identification of participants, and indication of the media and other files associated to the unit of recording.

The terms *metadata* and *annotations* have been used with some variation depending on the field of study, but a unified terminology is important for an integrated perspective such as the one presented here. The field of speech and language studies tends to refer systematically to *annotations*, although speech transcript (manual or automatic) is not always explicitly considered as an annotation of the speech signal. Image and video processing tend to call *metadata* the features extracted from the signals, which, in the case of still images, is still coherent with the definition adopted here. In the MPEG-7 proposals, the more generic term *descriptor* conflates the two types that are distinguished here [21].

11.2.2 Examples of large multimodal collections

This chapter draws from recent work involving the creation of large amounts of multimodal data accompanied by rich annotations in several modalities, and will often quote examples from recent projects dealing with the processing of multimodal data, such as AMI/AMIDA, CHIL, M4 and IM2 projects[1].

Most examples will be drawn from the AMI Meeting Corpus [8, 7], which is one of the most recent achievements in the field of large multimodal annotated corpora, together with the smaller and less annotated CHIL Audiovisual Corpus [24]. The AMI and CHIL corpora build on experience with data from previous projects, such as the ISL Meeting Corpus [6], the speech-based ICSI Meeting Recorder corpus [16], or the M4 Corpus [22].

Recording of multimodal data has, of course, started much earlier than the above-mentioned projects, but the resulting corpora (if constituted as such) are often smaller and lack the annotations that constitute the genuine value of data resulting from the projects quoted here. Other recent multimedia initiatives focus less on annotated data, and therefore have less challenges to solve for annotation: for instance, the TRECVID collection is used to evaluate the capacity to identify a limited number of *concepts* in broadcast news (audio-video signal), from a limited list, but the reference data includes no other annotation or metadata [31]. Many more contributions can be added if one counts also the work on multimodal interfaces, as for instance in the Smartkom project [33]. However, data management is a less prominent issue in the field of multimodal HCI, as data-driven research methods seem to be less used, at least until now.

[1] AMI/AMIDA EU integrated projects (Augmented Multiparty Interaction with Distance Access): http://www.amiproject.org, CHIL EU integrated project (Computers in the Human Interaction Loop): http://chil.server.de, M4 EU project (Multimodal Meeting Manager): http://www.dcs.shef.ac.uk/spandh/projects/m4/, IM2 Swiss National Center of Competence in Research (Interactive Multimodal Information Management): http://www.im2.ch.

11.3 Capturing and recording multimodal data

11.3.1 Capture devices

The capture of multimodal corpora requires complex settings such as instrumented lecture and meeting rooms, containing capture devices for each of the modalities that are intended to be recorded, but also, most challengingly, requiring hardware and software for digitizing and synchronizing the acquired signals. The resolution of the capture devices – mainly cameras and microphones – has a determining influence on the quality of the resulting corpus, along with apparently more trivial factors such as the position of these devices and the environment of the room (lighting conditions, reverberation, or position of speakers).

The number of devices is also important: a larger number provides more information to help defining the ground truth for a given annotation dimension. Subsequently, this annotation can serve to score signal processing over data from a subset of devices only, in order to assess processing performance over "degraded" signals. For instance, speech capture from close-talking microphones provides a signal that can be transcribed with better accuracy than a signal from a table-top microphone, but automatic speech recognition over the latter signal is a more realistic challenge, as in many situations people would not use headset microphones in meetings.

In addition to cameras and microphones, potentially any other sensor can be used to capture data for a multimodal corpus, though lack of standardization means that fewer researchers will be able to work with those signals. For instance, the Anoto® technology captures handwritten notes (as timed graphical objects), while eBeam® is a similar solution for whiteboards. Presentations made during recording sessions can be recorded for instance using screen-capture devices connected to video projectors, as in the Klewel lecture acquisition system (see http://www.klewel.ch). A large number of biological sensors can capture various states of the users, from fingerprints to heart rate, eye movement, or EEG. Their uses remain however highly experimental, since the captured data is often not general enough to be largely shared.

11.3.2 Synchronization

Synchronization of the signals is a crucial feature of a truly multimodal corpus, as this information conditions the possibility of all future multimodal studies using the corpus. Of course, the temporal precision of this synchronization can vary quite a lot, the best possible value being the sampling rate of the digital signals.

Although a primitive form of synchronization can be achieved simply by timing the beginning of recordings in each modality, there is no guarantee that the signal will remain time-aligned during the session, e.g. for one hour or more. Therefore, a common timing device is generally used to insert periodically the same synchronization signal in all captured modalities. For illustration purposes, this can be compared to filming the same clock on several video signals, but in reality the digital output of the synchronization device – such as a Motu Timepiece® producing a MIDI Time Code – is embedded in each of the signals, and most accurately in each sample of a digitized signal.

The synchronization accuracy is thus a defining feature of a multimodal corpus, and signals that are included in a corpus but with a lower synchronization accuracy face the risk to be ignored in subsequent uses of the data. One could argue that multimodal corpora lacking suitable synchronization information should better be called *plurimodal* rather than truly multimodal. Many of the more exotic capture devices raise difficulties in inserting synchronization signals in their data, as is the case with the Anoto pen technology, which first stores recorded signals in a proprietary format on the pen itself.

11.3.3 Activity types in multimodal corpora

Recording multimodal data collections requires a precise specification of the actions performed by the human subjects. In this chapter, we refer mainly to corpora of recorded human interactions, which can range from highly constrained settings in which participants behave more or less following a well-defined scenario, to highly natural ones, in which participants interact as if no recordings were taking place (ideally even unaware of the capture devices). Each of these approaches has its challenges in terms of finding subjects and setting the stage, as well as privacy issues; however, the setting fundamentally influences the future usability of the data. Therefore, a compromise between these constraints is generally found by the corpus creators.

Multimodal corpora are also collected in settings that do not involve human interaction, such as data for multimodal biometric authentication [29], in which case naturalness of (high level) behavior is often irrelevant. Such databases differ significantly from the ones considered here, especially in their temporal dimension.

11.3.4 Examples of setups and raw data

Gathering a large number of capture devices in a single place, with the purpose of capturing and recording multimodal data, has given rise to *instrumented meeting or lecture rooms*, also called *smart rooms*. Several large corpora were recorded in such spaces, in many cases using several physical locations set up with similar technical specifications. For instance, the AMI smart meeting rooms [23] were duplicated in three locations, as were the CHIL lecture and meeting rooms [13]. The NIST meeting room [32] was designed to acquire data mostly for the Rich Transcription evaluations. An example of recording procedure, for the Idiap meeting room used for the AMI and IM2 projects, is described in [15, pages 1–6].

The AMI Meeting Corpus includes captures from close-talking microphones (headset and lapel), from a far-field microphone array, close-up and room-view video cameras, and output from a slide projector and an electronic whiteboard. During the meetings, the participants also used weakly synchronized Logitech Anoto pens that recorded what they write on paper. All additional documents, such as emails and presentations, produced by the participants in series of meetings were collected and added to the corpus. The meetings were recorded in English using the three AMI rooms and include mostly non-native speakers.

The AMI corpus[2] consists of 100 hours of recordings, from 171 meetings. Most of the meetings (138 meetings, ca. 72 hours) are scenario-based, made of series of four meetings in which a group undertakes a design task for a remote control prototype [8, 7]. The remainder of the meetings (33 meetings, ca. 28 hours) are non-scenarised ones, mainly involving scientific discussions in groups of 3–5 people. As a comparison, the CHIL corpus comprises 46 lectures and 40 meetings, for a total length of about 60 hours, while the ICSI meeting corpus (audio only) consists of 75 one-hour recordings of naturally-occurring staff meetings.

Even if recording such large quantities of data has specific difficulties – like equipping smart rooms and eliciting data from groups of human subjects – the most costly challenge becomes visible only after the raw data was recorded: annotating a sizable part of the data. The usefulness of a corpus is not only determined by the amount of recorded data, but especially by the amount of annotated data.

11.4 Reference metadata and annotations

In this section, we discuss the process of encoding metadata and annotations once the raw data files have been captured and stored. Given our distinction between metadata and annotations, it is clear that metadata should take much less time to encode than the reference annotations, which, being time-dependent, are potentially also very time-consuming for human annotators (e.g. speech transcription takes between 10 and 30 times real time). While many tools and formats already exist for annotating multimodal data, the normalization of metadata is a much less explored topic, and as a result metadata for multimodal corpora is seldom encoded explicitly.

11.4.1 Gathering metadata: methods

Collection of metadata is often done at the time of recording, and despite the importance of this relatively small amount of information, the collection is often non-systematic and follows ad-hoc procedures that lead to incomplete records, which are nearly impossible to recover at a later stage. This is probably due to the fact that this information is heterogeneous in nature, and seems quite minor for each meeting considered individually. It is only when the collection of multimodal recordings is gathered that the role of the metadata in organizing the collection becomes obvious. These remarks apply of course to the *content* of the metadata information – its specific format can always be changed a posteriori.

[2] Publicly available from http://corpus.amiproject.org.

It is important for the future usability of a collection, and for integrating it in larger pools of resources (see Section 11.5.2), that metadata is gathered and consolidated into declarative files, with a content that is as detailed as possible, and a format that is as easily interpretable as possible. It is recommendable to follow existing and principled guidelines (also called norms or standards) for metadata encoding, to avoid developing a metadata specification from scratch. Well designed guidelines minimize the number of mandatory fields, and allow for user-defined fields, where information not foreseen in the guidelines can be encoded. It is not our purpose here to recommend particular guidelines, but some important examples are quoted below.

- MPEG-7 provides considerable detail for declaring low-level properties of the media files [21];
- Dublin Core was extended by the Open Language Archives Community (OLAC) to provide description entries mainly for speech-based and text-based data, with a relatively small number of descriptors, which can nevertheless be extended using specifiers [2];
- NXT, the NITE XML Toolkit [11, 9], handles implicitly part of the metadata with the annotation files (see next section);
- the IMDI guidelines [39, 3], designed by the Isle MetaData Initiative, are intended to describe multimedia recordings of dialogues [3]. The format offers a rich metadata structure: it provides a flexible and extensive schema to store the defined metadata either in specific IMDI elements or as additional key/value pairs. The metadata encoded in IMDI files can be explored with the BC Browser [4], a tool that is freely available and has useful features such as search and metadata input or editing.

11.4.2 Metadata for the AMI Corpus

For the AMI Corpus, the metadata that was collected includes the date, time and place of recording, and the names of participants (later anonymized using codes), accompanied by detailed sociolinguistic information for each participant: age, gender, knowledge of English and other languages, etc. Participant-related information was entered on paper forms and was encoded later into an ad-hoc XML file. However, the other bits of information are spread in many places: they can be found as attributes of a meeting in the NXT annotation files (e.g. start time), or in NXT "metadata" files, or they are encoded in the media file names (e.g. audio channel, camera), following quite complex but well documented naming conventions.

Another important type of metadata are the informations that connect a session, including participants, to media files such as audio and video channels, and the ones that relate meetings and documents. The former type is encoded in XML files

[3] See also `http://www.mpi.nl/IMDI/tools`

and in file names, while the latter is only loosely encoded in the folder structure. In addition, pointers to annotation files related to session, accompanied by brief descriptions of the annotations, should also be included in the metadata[4].

In order to improve accessibility to the AMI Corpus, the AMI metadata was gathered into declarative, complete files in the IMDI format [28]. Pointers to media files in each session were gathered from different XML resource files: mainly the `meetings.xml` and `participants.xml` files provided with the corpus. An additional problem with reconstructing such relations (e.g. finding the files related to a specific participant) was that information about the media resources had to be obtained directly from the AMI Corpus distribution web site, since the names of media resources are not listed explicitly in the annotation files. This implies using different strategies to extract the metadata: for example, stylesheets are the best option to deal with the above-mentioned XML files, while an HTTP-enabled "crawler" is used to scan the distribution site. In addition, a number of files had to be created manually in order to organize the metadata files using IMDI corpus nodes, which form the skeleton of the corpus metadata structure and allow its browsing with the BC Browser.

The application of the metadata extraction tools described generated the explicit metadata for the AMI Corpus, consisting of 171 automatically generated IMDI files (one per meeting). The metadata is now available from the AMI distribution website, along with a demo access to the BC Browser over this data.

11.4.3 Reference annotations: procedure and tools

The availability of manual annotations greatly increases the value and interest of a multimodal corpus. Manual annotations provide a reference categorization of the behavior of human subjects (in the most general sense) according to a given taxonomy or classification, in one modality or across several ones. These annotations are typically used for (a) behavior analysis with descriptive statistics, helping to better understand communicative behavior in a given dimension; and (b) for training and testing signal processing software for that respective dimension[5].

Reference annotations are done by human judges, using annotation tools which may include automatic processing. Annotations can be completed long after the data itself was captured, and can be made by several teams. In layered approaches such as NXT, certain annotations are based upon other ones, a fact that sets constraints on the execution order, and requires proper tools. Providing automatic annotations of a phenomenon has also potential utility, either as a sample of the state-of-the-art, or when manual annotation is not feasible for a large amount of

[4] Note that there are very few attempts to normalize the descriptors of the annotation files. Recent efforts to propose a unified description language for linguistic annotations were made by the ISO TC37/SC4 group [5].

[5] Strictly speaking, the terms *training* and *testing* are mainly used for systems that are capable of statistical learning; but in fact any type of software will require data for development or optimization, as well as reference data for evaluation.

data. The nature and resolution of the initial capture devices set conditions on what phenomena can be annotated, and with what precision.

Annotations have been done since data-driven methods and quantitative testing were first used, and therefore best practice principles are at least implicitly known for each modality and phenomenon. Such principles are related to the specific definition of the phenomenon to be annotated, the training of human annotators, the design of tools, the measure of annotators' reliability (often as inter-coder agreement), and the final validation of the result (e.g. by adjudicating the output of several annotators). However, most of these annotation stages raise various scientific issues, and the needs for an *annotation science* have recently been restated in [20, pages 8–10] for speech and language corpora.

A very large number of dimensions have been annotated in the past on mono and multimodal corpora. To quote only a few, some frequent speech or language based annotations are speech transcript, segmentation into words, utterances, turns, or topical episodes, labeling of dialogue acts, and summaries; among video-based ones are gesture, posture, facial expression, or visual focus of attention; and among multimodal ones are emotion, dominance and interest-level. Many annotation tools are currently available: some are specific to a given modality, while others enable the annotation of multimodal signals. Many tools are configurable to enable annotation using a categorization or tag set provided by the organizers of the annotations. A valuable overview of speech, video and multimodal annotation tools appears in [12], with a focus on nine currently used tools. For annotating multimodal signals, among the most popular tools are NXT, the Observer®, ANVIL [17], or Exmaralda [30].

For the AMI Corpus, annotators followed dimension-specific guidelines and used mainly but not exclusively NXT to carry out their task, generating annotations in NXT format (or similar ones) for 16 dimensions [10, 11]. Taking advantage of the layered structure of NITE annotations, several of them are constructed on top of lower-level ones. Using the NXT approach makes layered annotations consistent along the corpus, but renders them more difficult to use without the NITE toolkit. Due to the duration of the annotation process, not all the AMI Corpus is annotated for all dimensions, but a core set of meetings is available with complete annotations. The 16 dimensions are: speech transcript, named entities, speech segments, topical episodes, dialogue acts, adjacency pairs; several types of summary information (abstractive and extractive summaries, participant specific ones, and links between them); and in modalities other than language, focus of attention, hand gesture, and head gesture. Three argumentation-related items are also partially annotated. More detailed quantitative information about each dimension is available with the corpus distribution at http://corpus.amiproject.org.

11.5 Data storage and access

This section discusses issues of data, metadata and annotation storage and distribution, starting with hints about exchange formats, then continuing with requirements about data servers, and concluding with a real-time client/server solution for accessing annotations and metadata.

11.5.1 Exchange formats for metadata and annotations

Reusability is an important requirement for a multimodal corpus, given the large costs involved when creating such a resource. Therefore, the file formats that are used should be as transparent as possible, a requirement that applies to media files as well as to annotations and metadata. For media files, several raw formats or lossless compression solutions are available, and choosing one of them is often constrained by the acquisition devices, or by the resolution and sampling requirements.

Metadata and annotation formats are also generally associated to the tools that helped to encode them. However, when using annotations, it is often the case that the original format must be parsed for input to one's own signal processing tools. Given the variety of annotation dimensions, few exchange formats that are independent of annotation tools have yet been proposed. For instance, the variety of annotation formats for the language modality is visible in the MATE report [18], and more recent efforts within the ISO TC37/SC4 group have aimed at the standardization of language-based and multimodal dialogue annotation [5], through a repertoire of normalized data categories[6].

A solution for converting the annotations of the AMI Corpus into other usable formats was put forward in [28]. The initial goal of this solution was to convert the NXT XML files into a tabular representation that could be used to populate a relational database, to which browsing tools could connect (see Section 11.5.3). This conversion process can be easily modified to convert XML annotations to simpler file representation that can be used more easily by other software.

The conversion of the AMI NXT annotations proceeds as follows. For each type of annotation, associated to a channel or to a meeting, an XSLT stylesheet converts each NXT XML file into a tab-separated file, possibly using information from one or more other annotations and from the metadata files. The main goal is to resolve the NXT pointers, by including redundant information into the tables, so that each tabular annotation file is autonomous, in order to speed up queries to a future database by avoiding frequent joins. Upon batch conversion, an SQL script is also created, to load the data from the tab-separated files into a relational database of annotations. This conversion process is automated and can be repeated at will,

[6] Note that W3C's EMMA markup language – Extensible MultiModal Annotation – recently promoted as W3C Recommendation, is not a corpus annotation standard, but helps to convey content, in a multimodal dialogue system, between various processors of user input and the interaction manager.

in particular if the NXT source files are updated or the tabular representation must be changed.

For metadata, a similar process was defined, in order to convert the IMDI files – gathered from scattered or implicit information as explained in Section 11.4.2 – to a tabular format ready for a relational database. XSLT stylesheets were defined for this conversion, and the script applying them also generates an SQL loading script. Again, the stylesheets can be adapted as needed to generate various table formats.

11.5.2 Data servers

Multimodal corpora being collections of data/metadata/annotation files, they can be distributed as any set of files, using any digital support that appears to be convenient: CDROM, DVD, or even hard disks shipped from users to provider and back to users. Of course, these files can also be distributed via a network, from a file server, which can be the main storage server or not. Another approach to accessing these resources is via automated procedures that access one item at the time, on demand, as in a search application. These approaches are briefly discussed in this section.

An example of storage and distribution server for multimodal corpora is the MultiModal Media file server, MMM, based at the Idiap Research Institute, and hosting data for the AMI, IM2, M4 and other projects. The physical architecture of the server is described in [14], while the upload/download procedures are explained in [15]. The server offers secure storage for about 1 TB of data (the majority of the space is occupied by the video and audio files) and an interface that simplifies download by building for each request a multi-file download script that is executed on the user's side. For larger sets of media files, users must send hard disks that are shipped back to them with the data. The MMM server also manages personalized access rights for the different corpora.

In some cases, software components that use multimodal annotations and metadata do not need batch annotation files of past data, but rather need to query those annotations to retrieve specific items satisfying a set of conditions. These items can concern past data, but this case can be generalized to include situations when annotations produced by some signal processing modules are immediately re-used by other modules. Four examples of tools that allow specific querying over data/metadata/annotations are particularly relevant here.

1. A solution for accessing metadata from search interfaces was put forward by the Open Archives Initiative (see http://www.openarchives.org). In this view, corpus creators are encouraged to share their metadata, which is harvested via the OAI-PMH protocol by service providers which act as consolidated metadata repositories, and facilitate the discovery of resources regardless of their exact location.

2. The NXT system provides tailored access to annotations via the NITE Query Language [9].

3. A solution for accessing specific annotation items upon request has been developed, in relation to the AMI Corpus, as the Hub client/server architecture [1]. The Hub is a subscription-based mechanism for real-time exchange of annotations, which allows the connection of heterogeneous software modules. Data circulating through the Hub is formatted as timed triples (time, object, attribute, value), and is also stored in a database. *Producers* of annotations send triples to the Hub, which are received by the *consumers* that subscribed to the respective types. Consumers can also query the Hub for past annotations and metadata about meetings. At present, the entire AMI Corpus was converted to timed triples and inserted into the Hub's database as past data.

4. A solution for multi-channel and multi-target streaming was also developed to complement the Hub's mechanism in the domain of video and audio media files. The HMI Media Server (University of Twente, *unpublished*) can broadcast audio and video that is captured in an instrumented meeting room to various consumers, possibly remote, thus allowing a flexible design of interfaces that render media streams.

11.5.3 Accessing annotated multimodal data

Accessing multimodal data from an end-user's point of view involves the use of multimodal browsers and other tools that are capable of rendering media files and enhancing them with information derived from metadata and annotations. This section briefly introduces the concept of a meeting browser and exemplifies an application for speech-based retrieval of multimodal processed data.

The concept of meeting browsing has emerged in the past decade [25, 37], partly in relation to the improvements brought to multimodal signal processing technology. Meeting browsers generally take advantage of multimodal processing to improve users' access to meetings. This can mean either speeding up the search for a specific piece of information, or facilitating the understanding of a whole meeting through some form of summarization.

The JFerret framework offers a customizable set of plugins or building blocks which can be hierarchically combined to design a meeting browser. Using JFerret and other standard tools, several meeting browsers have been implemented. The creators of JFerret proposed sample instantiations of the framework, mainly based on speech, video, speaker segmentation, and slides [34, 35]. The JFerret framework was also used for other browsers, such as audio-based browsers, or dialogue-based, document-centric or multimodal ones in the IM2 project [19].

The Automatic Content Linking Device (ACLD) [27] demonstrates the concept of real-time access to a multimodal meeting corpus, by providing "just-in-time" or "query-free" access to potentially relevant documents or fragments of recorded meetings, based on speech from an ongoing discussion. The results are shown individually to meeting participants, which can examine the documents or the past meeting fragments in further detail, e.g. using a meeting browser, if they find that their contents are relevant to the current discussion.

Evaluation of access tools to multimodal data is a challenging topic, and evaluation resources such as BET [36, 26], which consist of true/false statements about a meeting produced by independent observers, could be added to the multimodal corpora as an extended form of annotation.

11.6 Conclusions and Perspectives

The volume of multimodal data that is recorded and made available as a collection is bound to increase considerably in the near future, due to the rapid diffusion of recording technology for personal, corporate or public use, and due to the improvement of processing tools that are available – although the robust extraction of semantic primitives from multimodal signals remains a major issue.

Most of the important problems that remain to be solved concerning data, metadata, and annotations are related to the challenge of increasing the interoperability of multimodal resources. The solution will involve the specification of shareable annotations, thanks to common data categories that can be further specified according to each project. Another challenge is the exchange of metadata [38] to facilitate the discovery of such resources via integrated catalogs.

This chapter has summarized the main stages of multimodal data management: design, capture, annotation, storage, and access. Each stage must be properly handled so that the result constitutes a reusable resource with a potential impact on research in multimodal signal processing. Indeed, as shown in the other chapters of this book, the availability of annotated multimodal collections tends to determine the nature and extent of data-driven processing solutions that are studied.

Acknowledgments

The author would like to acknowledge the support of the AMIDA and IM2 projects (see note 1).