Digitising Human Communications

The CAVA (human Communication: an Audio-Visual Archive) project will establish a digital video repository for human communication sciences, initially populated with a minimum of 600 hours of rights-cleared digital content owned by UCL research. CAVA project officer **Matt Mahon** provides an introduction to this new project.

In order to investigate human communication and interaction, researchers need hours of audio-visual data, sometimes recorded over periods of months or years. The process of collecting, cataloguing and transcribing such valuable data is time-consuming and expensive. Once it is collected and ready to use, it makes sense to get the maximum value from it by reusing it and sharing it among the research community.

But unlike highly-controlled experimental data, natural audio-visual data tends to defy easy classification, and may lead to idiosyncratic solutions to preservation, metadata and access issues. It is not uncommon for vital and unique data to languish on VHS tapes in personal collections. Natural data can often be used for more than the purpose its collector intended. Researchers may be able to save time and money, or improve the depth of their observations and conclusions, by reusing existing data instead of collecting their own.

Despite its usefulness, data in personal collections does not lend itself to being shared between researchers and institutions on a large scale. This is largely due to the absence, until now, of a centralised data archive to support such research and to offer opportunities for collaborative work. CAVA (human Communication: an Audio-Visual Archive) is a JISC-funded project based at University College London in collaboration with the UK Data Archive (UKDA), running from April 2009 to March 2010.

History and aims

CAVA was the product of a UCL Research Challenges grant which began in November 2007. This allowed the team to investigate the feasibility of centrally archiving data held by the Centre for Applied Interaction Research (CAIR), an interdisciplinary grouping largely based in the UCL Division of Psychology and Language Sciences. The CAIR project researched a discipline-specific metadata standard, and archived a pilot sample of data for dissemination through UCL's Moodle virtual learning environment. A study conducted as part of the CAIR project also found considerable support among the research community for a comprehensive and accessible repository.

CAVA will create and populate a repository which is accessible internationally by genuine researchers,

(© Asif Akbar)

and investigate the feasibility of having external researchers deposit data. The team will also work with the UKDA to investigate managed long-term storage of master files. Researchers in the CAIR group already hold a large amount of data with appropriate permissions for inclusion in the repository. By the end of the project the repository will hold 600 hours of data in various dissemination formats. There are three key areas to the project: Preservation, dissemination and access management.

Preservation

An important goal of CAVA is to centralise data held in personal collections and make it easily searchable by uniting the disparate sets of information held by the original researcher. The initial dataset will come from CAIR members based at UCL, though researchers at other institutions have expressed interest in contributing. The majority of this data is stored on portable media or cassettes, which run using proprietary or hard-to-find codecs, meaning that they are difficult to share and are at risk of obsolescence. The data will be captured from its original medium, whether CD, VHS or Mini-DV cassette, and transcoded into low-compression AVI format for preservation. The preservation file will maintain the quality of the original to the highest possible extent. This represents an acceptable compromise between fully uncompressed, but unwieldy, preservation files and the risk of the original formats becoming obsolete. The objective at this point is to achieve uniformity of data in order to aid long-term management of the dataset.

Dissemination

It is not enough, though, to just collect and standardise the quality of the data; it must be readily searchable. CAVA uses a modified metadata standard based on the ISLE MetaData Initiative (IMDI), a schema designed for language resources. The nature of the data presents some crucial challenges to the creation of metadata. Implementing the full IMDI standard would be too time-consuming and costly, for both the project team and depositors. The key issue to address is that of multiple participants. Based on various modifications of the IMDI standard, principally the UCL Deafness, Cognition and Language Research unit subset, the CAVA subset presents a pragmatic solution. A mapping between UCL's IMDI instance and the Dublin Core standard will be written. A full metadata schema and best-practice guides for capturing data (for both users and potential depositors) can be found on the CAVA website (www.ucl.ac.uk/ls/cava/).

The question of the dissemination formats has to be considered in tandem with the question of access and permissions. As the data is collected it will be stored using the UCL Library Services Digital Collections service, which runs on the Ex Libris DigiTool platform.

Access management

In order for a researcher to benefit from access to the data, they must be able to manipulate the files at their leisure. However, in order to encourage researchers to use the archive, and primarily to request access, the dissemination will operate on a tiered basis. Our key concerns at this stage are to ensure good procedures for data protection, identity and ethical issues. The metadata will be searchable through the DigiTool front page, although none of the data itself will be viewable at this point. A bespoke login will allow researchers to view streamed sample videos of the data which interests them.

> … an important goal of CAVA is to centralise data held in personal collections and make it easily searchable

The team will devise and test ingest processes so that video clips, transcripts (where available) and descriptive metadata can be uploaded to the repository in batches, in a way which maintains the relationships between the one or more versions of each video recording, its transcript, and the metadata which applies to each. The final ingest process will include the automatic generation of technical metadata and the creation of appropriate access restrictions.

(© Sanja Gjenero)

The data will be made available in several dissemination formats. All data accepted by the archive will have appropriate permissions for the various types of dissemination. Users will be available to download compressed video or uncompressed audio-only files. All dissemination formats will be prepared in order to operate on managed computers with minimal codecs and system requirements. A guide to format specifications can be found on the CAVA website.

The researcher would then request access to downloadable versions of their selected datasets. By these means, CAVA takes all reasonable precautions to prevent the often-sensitive data from being used inappropriately.

The project team will work with the UKDA to design the application process for prospective users, implement procedures to verify and authorise requests, and register and authenticate users. The retention and presentation of rights information will be implemented within the IMDI record. Users will see hard-copy or click-through licensing agreements to be associated with particular tiers of access, to indicate clearly and unambiguously what an authorised user may and may not do with the material.

Conclusion

The two key challenges faced in the delivery of the archive are those of appropriate metadata and good management of rights and access. Through extensive planning and preparation the team is actively ensuring not only that the repository is useful for the human communication research community, but also that the data we manage is secure and safe.

Matt Mahon
Project Officer (CAVA)
UCL Division of Psychology and Language Sciences
Email: lib-cava@ucl.ac.uk
www.ucl.ac.uk/ls/cava