



JISC Final Report

Project Information			
Project Acronym	CAVA		
Project Title	Human Communication: an Audio-Visual Archive for UCL		
Start Date	01 April 2009	End Date	31 August 2010
Lead Institution	UCL		
Project Director	Paul Ayris		
Project Manager & contact details	Martin Moyle m.moyle@ucl.ac.uk / 020 7679 4351		
Partner Institutions	UK Data Archive		
Project Web URL	http://www.ucl.ac.uk/ls/cava		
Programme Name (and number)	Capital Programme - Repositories: start-up		
Programme Manager	Amber Thomas; Andrew McGregor		

Document Name			
Document Title	Final report		
Reporting Period			
Author(s) & project role	Martin Moyle, Principal Investigator Dr Suzanne Beeke, Co-investigator Matt Mahon, Research Assistant Dr Merle Mahon, Co-investigator		
Date	01 October 2010	Filename	
URL	http://discovery.ucl.ac.uk/575242		
Access	Public		

CAVA

A Human Communication Audio-Visual Archive for UCL

Final project report

Martin Moyle, UCL Library Services

Dr. Suzanne Beeke, UCL Language and Communication

Matt Mahon, UCL Language and Communication

Dr. Merle Mahon, UCL Developmental Science

October 2010



Contents

Acknowledgements	4
Executive summary	5
Background	6
Aims and objectives	6
Methodology	7
Implementation	8
Outputs and results	10
Outcomes	17
Conclusions	18
Implications	18
References	19

Acknowledgements

The CAVA project was part-funded by the JISC (Joint Information Systems Committee), under the JISC Capital Programme - Repositories: start-up strand, with matching funding from UCL.

The project was a collaboration between Research Departments from UCL's Division of Psychology and Language Sciences, UCL Library Services, and the UK Data Archive.

Project Team

- Dr Suzanne Beeke, Head of Department, UCL Language and Communication
- Matt Mahon, UCL Language and Communication, Project Officer
- Dr Merle Mahon, Senior Lecturer, UCL Developmental Science
- Martin Moyle, Digital Curation Manager, UCL Library Services, Project Manager

Steering Group

- Dr Paul Ayris (Chair), Director of UCL Library Services and UCL Copyright Officer
- Dr Suzanne Beeke, Head of Department, UCL Language and Communication
- Dr Libby Bishop, Research Liaison, ESDS Qualidata, UK Data Archive and Senior Research Archivist, Timescapes, University of Leeds
- Dr Merle Mahon, Senior Lecturer, UCL Developmental Science
- Martin Moyle, Digital Curation Manager, UCL Library Services
- Stevie Russell, Site Librarian, UCL Language and Speech Science Library

Thanks also to the members of the UCL Centre for Applied Interaction Research (CAIR) who gave up their time to provide useful feedback on the repository in the course of the project, and to the [Timescapes](#) project team.

Executive Summary

The objective of the CAVA project was to establish a repository for primary audio-visual data on real-life human communication for spoken and signed languages.

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 has been collected, its value to the research community can be maximised by re-use. The CAVA repository was designed to store, and make available to bona fide researchers worldwide, audio-visual recordings collected by members of UCL's Centre for Advanced Interaction Research.

The approach taken was to customise the UCL Library Services Digital Collections service to enable it to harbour a dedicated audio-visual service. The issues addressed included file formats, both for dissemination and long-term preservation, and format conversion; metadata, notably the creation of a schema for multimedia human communication resources, derived from the IMDI (Isle Metadata Initiative) standard; the user interface; the acquisition of content; procedures and processes in support of rights management and access restrictions; guidance and training materials, both for prospective depositors and users; the need for hands-on mediation of the repository deposit process; dissemination, not only about the work of the project, but also to attract content into the repository and to notify potential users of its availability; and sustainability, both in terms of content and services.

The issues were addressed successfully and the project has delivered a repository service to the human communication research community, housing some 750 hours of primary video material, supported by custom metadata and underpinned by robust access management.. Some questions of sustainability remain to be addressed, largely around the availability of resources to expand the content of the repository, the remit of the repository, or both; but the project team is committed to sustaining the basic service with existing levels of content. Digital preservation issues have been explored with the support of the UK Data Archive, a CAVA project partner, and an innovative 'preservation partnership' between UCL and the UK Data Archive has been agreed in principle.

All the project documents, reports and training materials produced by CAVA are available through the CAVA website.

1. Background

The objective of the CAVA project was to establish a repository for primary audio-visual data on real-life human communication for spoken and signed languages.

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 has been collected, its value to the research community can be maximised by re-use.

Historically, the study of communication has been based on highly-controlled experimental data, but a better understanding comes from examining audio-visual data of natural behaviour. Such work, both qualitative and quantitative, involves in-depth study of video- and audio-recorded data of conversations and clinical encounters. However, unlike highly-controlled experimental data, natural audio-visual data tends to defy easy classification, and this may lead to idiosyncratic solutions to preservation, metadata and access issues. It is not uncommon for unique data to languish on VHS tapes in personal collections: researchers across the discipline waste time battling with increasingly inaccessible media and finding individual solutions to the challenges of editing and analysis. The resources of funders, researchers and subjects are inefficiently expended on the collection of new data rather than on the re-use of existing data, which might easily be applicable to new researchtopics. The CAVA project was designed to support the premise that researchers in human communication might be able to save time and money and improve the depth of their observations and conclusions by reusing existing data.

The work of CAVA began in the shape of a UCL Research Challenges grant in 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 investigated a discipline-specific metadata standard, and archived a pilot sample of data for dissemination through UCL's Moodle virtual learning environment. A feasibility study conducted as part of the CAIR project also found considerable support among the research community for a comprehensive and accessible repository. This work was the precursor of CAVA, which created a digital repository of rights-cleared, re-usable audio-visual material to support the work of UCL and the international human communication research community.

The CAVA partners were three UCL Departments, UCL Library Services, UCL Developmental Science and UCL Language and Communication, and the UK Data Archive.

2. Aims and objectives

The CAVA repository was to be housed within the UCL Library Services Digital Collections Service, which is based on the Ex Libris DigiTool platform. A discipline-specific metadata standard, IMDI (ISLE Meta Data Initiative) was to be adopted. The material for deposit in the repository involves human subjects in conversation, and for ethical reasons it cannot be made openly accessible. Access management procedures to enable the authorisation and authentication of bona fide researchers were therefore to be devised. Finally, digital preservation of the CAVA content was to be investigated.

CAVA's main aims, in summary, were as follows:

- To configure the CAVA repository as a discrete archive within the UCL Digital Collections service.
- To implement support for the IMDI metadata standard to an appropriate level of detail, and to ensure metadata interoperability between the CAVA repository and Dublin Core-based aggregation services.
- To populate the CAVA repository with a minimum of 600 hours of rights-cleared audio and video material in agreed dissemination formats, accompanied by transcripts and other supporting material where available.
- To ensure that the ongoing population of the CAVA repository is embedded within the relevant UCL research communities.
- To implement processes and technical procedures to support the management of access to the content of the repository.
- To publicise the CAVA resource to potential users and other stakeholders.
- To investigate the feasibility of the expansion of the remit of the repository to encompass deposit by non-UCL researchers in appropriate disciplines.
- To appraise the options for the managed long-term storage of uncompressed master files and to support the digital preservation of the CAVA corpus.

An additional aim, to create training and guidance material tailored for the repository's two main audiences, depositors and users, was agreed in the course of the project.

3. Methodology

CAVA fell broadly into three lines of work: creating the repository, dissemination about the project to stakeholders, and ensuring sustainability. The project benefited in all areas from partnership with the UK Data Archive, whose experience and good practice were drawn into the planning wherever possible. In particular, early in the project the team undertook a learning visit to the UK Data Archive, whose staff kindly shared details of policy and practice in the areas of rights and access management and digital preservation.

The creation of the CAVA repository involved the customisation of an existing repository platform, DigiTool, with no new technical development work. The typical methodology for each stage of the customisation process was that a particular output was piloted or drafted and then validated, in a process conducted by the project team with the support of members of CAIR, colleagues from the UK Data Archive, or both. After successful validation, the output was modified accordingly and rolled out.

In its sustainability work, the team needed to address the sustainability both of the CAVA service and its underlying content. To address the latter, an options appraisal for digital preservation was undertaken, and a short report produced. With regard to the service, a sustainability discussion paper was drafted and discussed with key stakeholders, to help to inform the team's thinking on the issues of post-project service levels, costs and funding options. This was supplemented by an end-of-project survey of depositors and users, conducted online using SurveyMonkey.

In constructing a dissemination plan, the team recognised that the project had a range of potential stakeholders - among them depositors, researchers, teachers, students (especially PhD researchers), libraries and digital repository specialists, and research funders. Plans for project dissemination activity focused initially on general awareness-raising, then on the recruitment of content for the repository, and finally on the active promotion of the repository as a resource for users.

4. Implementation

The repository

The work of implementing the CAVA repository was divided into six areas: formats, metadata, interface, content, access and guidance.

4.1. File formats

The data which was initially identified for deposit in the CAVA repository was held by researchers in a range of formats, with an equally wide range of software requirements. An analysis of appropriate file formats for CAVA was undertaken with input from local UCL experts, and with validation by the UK Data Archive. It was recognised that these would have to be somewhat ‘aspirational’ for legacy content, but that, going forward, best practice format recommendations tending to uniformity might both help the management of the repository and help the decision-making of researchers creating new audio-visual content.

4.2. Metadata

Pre-project work had confirmed that the IMDI (ISLE Meta Data Initiative) schema, a standard for the description of multimedia and multi-modal language resources, was appropriate for CAVA. The main challenge was to find the right balance between detailed description, the burden on depositors when compiling metadata, and the discovery and documentation needs of the eventual users of the resource. A subset of the IMDI standard, with project-specific controlled vocabularies where appropriate, was drawn up, and validated by members of CAIR before implementation.

4.3. User interface

A CAVA interface to the UCL Library Services Digital Collections service was designed, tested by CAIR researchers, and finalised.

4.4. Content acquisition

Content acquisition involved advocacy, to persuade potential depositors of the benefits of re-use (summarised under Dissemination, section 4.9); format conversion, according to the format specifications drawn up by the project team; and upload.

In order to lower barriers to participation, project staff assisted researchers with metadata creation and carried out all necessary file conversions. Upload was also carried out by project staff. To this end, ingest processes in which audio and video files and transcripts (where available) and descriptive metadata could be uploaded to the repository in batches, with automatic technical metadata generation, and the maintenance of the relationships between the one or more versions of each video recording, its transcript, and associated metadata, were devised and tested.

4.5. Access and rights management

Rights and access issues formed a substantial part of the work of CAVA, because of the sensitivity of the video material. It was essential that proper consent had been secured before such material could be deposited; after deposit, open access to CAVA content is neither possible nor desirable. Bona fide researchers, however, should be able to access and re-use CAVA material without too many barriers. The team drew up several model documents to support rights management: a model consent form for use by CAIR researchers, a repository deposit licence, and an end-user licence for the repository. The team also explored the procedures and technical processes necessary to manage access to the content, taking into account the need for user application, authorisation, registration and subsequent authentication. The rights work incorporated a learning visit to the UK Data

Archive, to ensure close alignment between CAVA policy and the data policies of research funders, and validation by UCL's Data Protection Officer.

4.6. Guidance and instruction

It was recognised that training and guidance materials would be required, and that these were addressing two audiences, depositors of content and users of deposited material. Material for those in the former category needed to provide assistance from the very first stages of a project using Conversation Analysis (Schegloff 2007) procedures, when ethical and rights issues are addressed, as well as practical matters such as file formats and CAVA metadata. For users, guidance was required to be available in different formats, and had to address not only the background to CAVA and its content, and the registration process, but technical issues around downloading and rendering, in different browsers, files created with different codecs.

Sustainability

CAVA's sustainability work focused on two areas: the sustainability of the content - that is, its digital preservation - and the sustainability of the CAVA repository service.

4.7. Digital preservation

Two aspects of digital preservation were considered, day-to-day – encompassing file integrity, storage and back-up concerns - and long-term. With regard to the former, maintenance of the integrity of the CAVA dissemination files is already well supported within the repository environment: the use of redundancy and mirrored storage, off-line backups and checksums mean that the team can have reasonable confidence in the 'bit-level' preservation of the dissemination files housed within the repository.

Preserving these data in the long-term introduces some new challenges: what happens when file formats (or repository platforms) become obsolete? What measures need to be in place to ensure that we know when such obsolescence has occurred? Again, in some respect the team was able to address these issues. The file format recommendations (see 5.1) ensured that master files were created in non-proprietary, lossless formats, from which new derivatives could be created in future if necessary. The team ensured that the CAVA deposit licence gives the repository permission to translate files to new formats. UCL Library Services has a digital preservation policy; and technical metadata, essential to long-term preservation, is generated and stored about any file which is added to the CAVA repository.

The main issue for CAVA was long-term storage. Audio-visual file streams are large, and it was clear at the outset of the project that the preservation masters for CAVA deposits, an originally estimated 8 terabytes of data, would be too big to be housed within the Library's managed storage environment.

Three courses of action, not necessarily mutually exclusive, were considered.

(i) The use of off-line media – portable USB hard drives – for the long-term storage of master files. This is far from the ideal solution, but likely to provide temporary help nonetheless.

(ii) Participation in a large-scale UCL storage facility. Like most research-intensive, multi-disciplinary universities, UCL clearly has a corporate need to support the long-term storage and management of primary data on a large scale. However, no institution-wide data storage facility was likely to become available within the lifetime of the project.

(iii) Archival deposit with the UK Data Archive, or other trusted third party.

This is potentially an elegant solution. It was recognised that, if a willing partner were found, issues of logistics, policy and resources would have to be explored, for instance how and when would files be transferred to the archive, and what documentation and/or metadata

would need to accompany them? What about future deposits, should the repository continue to expand? What arrangements would need to be in place for the retrieval of CAVA masters from the archive, either wholesale? Careful scrutiny of policies and facilities would be required – are there precedents? What ‘leasing’ costs would fall to the repository; what other resources would need to be committed; and how would all these resource requirements be met?

4.8. Service sustainability

The UCL partners were committed to the maintenance of the CAVA repository beyond the lifetime of the funded project. The challenge when drawing up the project exit strategy was to find the right balance between the services to be offered in continuation and the availability of resources to deliver them. These discussions are summarised in section 5.8.

4.9. Dissemination

Dissemination activity began with awareness-raising within CAIR, as well as within the JISC and DigiTool communities. When the repository was established, a content recruitment drive took place within CAIR, and among other research groups with close links to the Centre. The CAIR community was included in the design of the metadata vocabularies, which had the effect of demonstrating the ease of completing the metadata.

Finally, the repository as a resource was promoted to potential users. Users were recruited from CAIR research seminars, the ICPLA 2010 international conference, the UCL TILT conference, and the Chandler House Festival (a Divisional open day). Invitations to join were circulated on the British Association of Clinical Linguistics (BACL), and British Association of Applied Linguistics (BAAL) mailing lists. CAVA details were posted to the Ethno/CA News site, where the repository remains a featured audiovisual resource.

5. Outputs and results

5.1. File formats

It was determined that each video deposited with CAVA should be held in up to three formats: preservation master, dissemination (video) and dissemination (audio). A copy in at least one dissemination format is mandatory, and a preservation master is highly desirable, formats permitting.

The dissemination files are served from the repository directly to the local PC. This enables a researcher to process and analyse files using their preferred interaction data methodology and/or software. A streaming service is not required. The team considered the possibility of creating small sample videos in a streaming format, to help researchers to explore the repository and identify datasets appropriate to their work before going through the registration process: this is a possible future enhancement to the service, although it was not implemented in the course of the project.

The team’s file format recommendations are summarised in the following table. The full report may be found at the CAVA website.

File type		Capture	Master	Download	Streaming (UCL standard)	Audio-only
		AVI	AVI	MPG	FLV	WAV
Video	Codec	[DVSD]	DV25	MPEG-1	On2 VP6	N/A
	Data rate (kbps)	28800	28800	3024	400	N/A
	Frames/sec	25	25	25	25	N/A
	Frame size	720x576	720x576	720x576	480x360	N/A
Audio	Codec	PCM	PCM	MP2	MP3	PCM
	Data rate (kbps)	1024	1024	224	128	1024
	Sampling rate (Hz)	44100	44100	44100	44100	44100
	Channels	2	2	2	2	2
	Sample precision	16-bit	16-bit	16-bit	16-bit	16-bit

Fig.1. Summary of CAVA file format recommendations.

Transcripts, where available, are accepted in pdf, MSWord or CHAT format (see <http://childe.psych.cmu.edu/>).

5.2. Metadata

The metadata fields used in the repository to describe CAVA content are shown below. Optional content is denoted by square brackets. (c) denotes that a controlled vocabulary is used.

No.	<i>Object +</i>	
1	Identifier	
2	Date (c)	
3	Original format (c)	
4	Format history	
	<i>Location (sub)</i>	
5		Country (c)
6		Description
	<i>Project +</i>	
7		Name
8		ID
		<i>Contact (sub)</i>
9		Name
10		Contact's organisation
11		Longitudinal project (boolean)
12		Description
	<i>Content +</i>	
13	Genre	

14		Subgenre
15		Communication Context
		<i>Languages</i> (sub)
16		Number of languages (c)
17		Spoken language ID (c)
18		Sign language ID (c)
19		Language variety
20		Communication modes
		<i>Transcription</i> (sub)
21		Transcription (boolean)
22		[Transcription format]
	<i>Actors</i> +	
23		ID
24		Age (c)
25		Age band (c)
26		Sex (c)
27		[Occupation or previous occupation]
28		[Actor notes]
		<i>Condition</i> (sub)
29		Condition
30		Condition subtype
31		Cause of condition
32		Onset of condition
33		Intervention history
34		Family history
35		[Hearing status]
36		[Vision status]
37		[Handedness]
38		[Sign language experience]
	<i>Education</i> (sub)	
39		[Education leaving age] (c)
40		[School Type]
41		[Class Kind]
42		[Education Model]
43		[Boarding School] (boolean)
44	Secondary actor(s) notes	
	<i>Access</i> +	
45		Rights (c)
46		Rights evaluation date (c)
47		Owner

Fig.2. Descriptive metadata fields used in CAVA repository

The CAVA standard is derived from the IMDI standard. A full report on CAVA metadata, with element descriptions and indicative vocabularies, is available from the CAVA website.

5.3. User interface

CAVA content is housed within the UCL Library Services Digital Collections service. A general search of Digital Collections will bring up CAVA content. However, a dedicated CAVA interface to the service, with CAVA branding, and through which only CAVA content may be searched and browsed, was created. The CAVA site has a direct URL, which CAVA repository users are encouraged to use.

UCL LIBRARY SERVICES
CAVA Repository

Library home > Online resources > Digital Collections > CAVA Repository

Search | Results | Search History | e-Shelf | Help | About

You are not logged in: Log in here

Search · Advanced search · Browse by project · Browse by condition

Enter search term(s):

Record: Contains Exact Starts With

CAVA Repository

The CAVA (human Communication: an Audio-Visual Archive) repository contains rights-cleared primary audio and video recordings made by UCL's human communication researchers and their associates.

To find data appropriate to your research, you can either:

[SEARCH](#) for a specific term, for example "*Sensor-neural bilateral hearing loss*", or
[BROWSE](#) the collection by project, condition, age or education.

In order to view any files, you will need to register. [Information on how to join can be found here](#). Please log in using the links below:

■ UCL users: [Log in here](#)
■ External users: [Log in here](#)

For more information, follow the links below:

■ A video tour of the archive, via our [YouTube channel](#):

Fig.3. The CAVA Repository - home page: <http://www.ucl.ac.uk/library/cava>

The full metadata description (see 5.2) and the full text of any uploaded transcripts are indexed for search. Information about projects, age of subject, and the communication disorder and/or medical condition of subjects is indexed for more focused searching, if required. The archive also offers drill-down browse navigation, by project or condition.

Results lists show brief metadata descriptions. Padlocks denote that login is required.

UCL LIBRARY SERVICES
CAVA Repository

UCL

Library home > Online resources > Digital Collections > CAVA Repository

Search | Results | Search History | e-Shelf | Help | About

You are not logged in: Log in here

Results

Search 'W-All Words= deafness or aphasia' in 'CAVA Repository' Collection [Sorted by: Condition]

Brief view Table view Full view

Sort by: Condition

Records 1- 20 of 507

1 **DIC1**
The evaluation of a novel conversation-focused therapy for agrammatism
11/08/2008
aphasia

2 **DIC2**
The evaluation of a novel conversation-focused therapy for agrammatism
18/08/2008
aphasia

3 **DIC3**
The evaluation of a novel conversation-focused therapy for agrammatism
26/08/2008
aphasia

4 **DIC4**
The evaluation of a novel conversation-focused therapy for agrammatism
01/09/2008
aphasia

5 **DIC5**
The evaluation of a novel conversation-focused therapy for agrammatism
09/09/2008
aphasia

6 **DIC6**
The evaluation of a novel conversation-focused therapy for agrammatism
14/10/2008
aphasia

Fig.4. The CAVA Repository - search results

Brief results may be expanded to show the full metadata record for a resource.

Record 2 of 43

Object

CHA

1 JOL 10-04 TA - CHATfile (16 K)
1 JOL 10-04 TA - Word Document (764 K)
1 JOL 10-04 TA - MPEG file (136 M)
1 JOL 10-04 TA - WAV File (101 M)

Resource title

1 JOL 10-04 TA

Project

EAL Deaf Children .

Date

2004-10

Project ID

HMM-DH

Project Contact

Dr Merle Mahon

Contact's Organisation

UCL

Project Description

Deaf children conversation

Longitudinal Project

Yes

Content - Genre

One:One

Content - Subgenre

Classroom, child and teacher

Communication Context

Booksharing

Actor ID

JOL

Age

3;6

Sex

Male

Actor Notes

L1 English

Condition

Deafness

Condition Subtype

Sensor-neural bilateral hearing loss

Cause of Condition

Respiratory distress at birth, 2 months NICU

Onset of condition

Congenital

Intervention History

diagnosis 0;2; bilateral digital hearing aids 0;3; cochlear implant right ear 6;1; preschool peripatetic teacher of the deaf + speech and language therapy; entered school 3;5; speech and

Fig.5. The CAVA Repository - full metadata record (excerpt). This particular resource is available in video (mpg) and audio-only (wav) versions; transcripts in MSWord and with CHAT encoding are also available.

Once a registered user logs in, audio-visual files are delivered to the desktop, for analysis using the researcher's preferred software application.

5.4. Content acquisition

By the end of the project, the repository held 2792 resources, representing some 750 hours of video. The dissemination files occupy 1.3 TB of storage space. Deposit was entirely mediated by the project research assistant.

5.5. Access and rights management

All CAVA data is subject to access restrictions. A registration process was established which is informed by the procedures developed for the Timescapes Archive and also aligned with practice at the UK Data Archive. Applicants download a user licence from the CAVA web site, stating who they are and their research need to access the data. The application is assessed by a member of the CAVA team. Once authorised, UCL users authenticate using their institutional username and password; external users are supplied with a 'guest' username and password. These usernames allow access to any of the items in the repository on which additional access restrictions have not been imposed by the depositor.

A policy on student access was developed. Undergraduate students are not eligible for CAVA membership, although the CAVA licence permits a researcher to show CAVA data in teaching situations. Postgraduate students are not eligible for membership unless their application is reviewed by one of CAVA's academic partners. This is because not all Masters-level work falls under the use category 'research' in the original permissions given by the participants in the recordings. Doctoral-level students undertaking relevant research may access the repository.

To help researchers secure the permissions necessary from subjects when collecting video data, pro forma consent forms for adult and child subjects were developed.

5.6. Guidance and instruction

Various outputs were prepared to help both depositors and data users. A video introduction to the repository was embedded in the home page, served from the CAVA YouTube channel. A FAQ and a technical troubleshooting guide were also made available from the repository home page. File format and metadata guidance, guidance for users citing CAVA resources, and guidance for researchers wishing to incorporate CAVA deposit in future funding applications were also made available through the CAVA repository website.

5.7. Digital preservation

Ideally, CAVA master files would be stored within an archival area of the UCL Library Services Digital Collections repository. As outlined at 4.7, because of limitations on the size of the managed storage available, this is not feasible. Consequently, the CAVA master files will be kept on portable storage, to be held in the custody of the Library. Meanwhile, the team is working with the UK Data Archive towards an agreement whereby preservation masters are held, in a managed environment, by the UK Data Archive, with fundamental preservation services also provided.

An agreement-based relationship with a data archive is felt to be the most responsible way for CAVA to tackle the outstanding preservation issues, provided that the cost and other issues are surmountable, at least until such a time as the host University can offer appropriate large-scale data storage facilities on which to base long-term digital curation.

5.8. Service sustainability

Post-project, the repository will be maintained by UCL Library Services, with administrative support and academic direction from CAIR. The requirement is to find the right balance between desirable levels of service and available funding. In order to determine how the repository should look beyond the funded lifetime of the project, the team considered different service levels, their associated costs, and the availability of resources.

The most fundamental level of service, and the minimum that CAVA will endeavour to support in the long-term, is to continue to serve out the content already acquired, but to close the repository to new deposits.

With sufficient resources, the repository will continue to expand. Basic expansion would involve adding new data from current and future human communication research projects at UCL. Further expansion possibilities might involve taking content from human communication research groups, outside UCL, whether based in the UK or overseas; and/or to begin to cater for other disciplines also requiring audio-visual repository services.

Clearly, in any expansion scenario, sufficient funds would have to be available to guarantee sustainability for the medium- to long-term.

In assessing the costs of different service levels, an overview of the responsibilities needed to maintain the repository may be helpful. The following roles and responsibilities would be necessary to fulfil a basic CAVA continuation service, offering access to existing content without any expansion of holdings:

- Registration and authorisation of new end-users
- Rights renewal and expiry
- Server-side authentication management
- Server management
- Application management
- Enquiries management
- Revisions to documentation and guidelines, e.g. in response to changes to repository platform, new playback and analysis software, new browsers
- Training and guidance for staff involved in CAVA support
- Library-academic liaison
- Digital preservation, with outsourcing payments if appropriate

The above tasks are fundamental to CAVA, and will apply at some scale whatever services are offered in future. Basic server and application management are part of UCL Library Services' wider commitment to its digital repositories; however, the maintenance of CAVA, with its additional demands in areas such as metadata, rights, interface and formats, is more resource-intensive than a basic repository service.

Were any commitment to be made to expanding the content of the CAVA repository, additional support would be required in the following areas:

- Support for metadata preparation
- Advice on and software support for file conversions
- Ingest
- Checking deposit licences
- Maintenance and development of user interface: building navigation across new content, any other changes to the user interface
- Hands-on guidance, training and support for depositors
- Publicity to designated users
- Support for users, e.g. on different codecs, new browsers and browser versions
- Storage expansion: procurement, deployment and management
- Commitments to best practice
 - Discussion and implementation of revisions to metadata schema
 - Implications of new file formats for conversions, storage, training and guidance materials, preservation, legacy content

- Refreshment of publicity and training materials, and their propagation in new media
- Management reporting (for instance, usage figures, to promote the repository or to support the case for new funding)

Finally, if expansion to disciplines other than human communication were brought into play, start-up costs that were relevant to CAVA, such as metadata analysis and implementation, would apply once more. A need for streaming services might also be anticipated.

Funding, over and above the commitment of UCL Library Services and the CAIR partners to maintaining the basic repository service with its existing content, is not currently available. The main post-project requirement is to identify secure long-term funding for CAVA, offering whatever services suit CAIR, UCL Library Services, and the funder or funders.

Possible funding sources include contributions of UCL staff time; applications to research funding agencies - grants to sustain the repository per se are unlikely to be available, but costing for CAVA use could be included in applications to research funding agencies; commercial partnership, sponsorship and/or advertising; and - perhaps more of a theoretical possibility - benefactors.

Realistically, it is almost inevitable that CAVA will depend on 'hand-to-mouth', piecemeal funding from this point onwards, and that the team will have to be both reactive and flexible. It is also noted that any short-term funding in return for new deliverables may create additional sustainability problems in the longer term, and therefore such opportunities should be treated with a degree of caution.

The team has concluded that research grants represent the most realistic source of continuation funding for CAVA. CAIR members will be encouraged to add a contribution to staff and storage for the CAVA repository in any future grant applications. To this end, some short guidance for CAIR members on how to explain the CAVA service and its merits to funders, and how to calculate and incorporate a CAVA component in the costings for a proposed project, has been drawn up. The policy of encouraging researchers to seek CAVA costs in applications explicitly ties repository expansion to the availability of funds to support it: new content will only arrive if grants are successful, and each grant will include a proportionate resource contribution for the repository. It is felt that this is the best way to ensure the sustainability of the repository in the longer term.

5.9. Dissemination

The main vehicle for CAVA dissemination was the project website, which houses all the publicly available project documents as well as project news. CAVA's own mailing list is used to disseminate news and updates about the service. Various conferences, seminars and professional mailing lists were used to raise awareness of the project and the service - these are listed at 4.9. In terms of publications, the team was able to present two conference posters, make two conference presentations, and publish a magazine article.

6. Outcomes

The core aims of the project were met. The methodology used to design and implement the repository was successful. The CAVA repository was configured as a discrete archive housed within the UCL Digital Collections service, underpinned by a discipline-specific metadata standard and best-practice access management procedures. The repository has been publicised widely within the research community. Sustainability plans for the repository are in hand.

The project has brought a number of benefits to the HE community. The repository is a significant new research resource available to all researchers in the human communication discipline. It has set a precedent for the repository-supported re-use of video data in the field, in a way that supports different analytical methodologies without constraint. The team has acquired and shared expertise in the managed curation of video materials which is relevant to other research communities, across disciplines, including material for which consent and rights present additional challenges.

7. Conclusions

The CAVA repository start-up project delivered an audio-visual repository of human communication research data holding over 750 hours of material. Research carried out before the project began had indicated that there was demand for such a service; the repository is now seeing evidence of use, with applications from as far afield as Dubai, Singapore and Australia.

The management of rights-protected video material brings a few technical challenges that are not commonly addressed at this scale by standard institutional repositories. CAVA's work encompassed file formats and conversions, the dissemination challenges of different codecs and their interaction with browsers, and the preservation challenges of video material, not least that of file size. In this respect, the project has developed an innovative partnership with a national data centre, the UK Data Archive.

The team demonstrated a model for community involvement in drawing up a metadata standard to suit a particular discipline and for implementing it within the framework of a general digital collections service. Although the CAVA metadata scheme was distilled from a much larger standard, focusing on the essentials to balance ease of use against breadth of coverage, it is still relatively complex, and for this reason, allied to the general complexities of file management in this setting, a mediated service to depositors was deemed to be essential. This naturally raises the cost of any repository operation, and was taken into account when CAVA's sustainability options were considered. Asking prospective depositors to write CAVA costs into research applications seems the most logical way to support the ingest of new content into the repository: guidance on preparing such costings has been drawn up for the benefit of UCL researchers.

The nature of the material deposited in CAVA demands an unusually (for a repository project) strong emphasis on rights and access management. In this respect, CAVA showed that much is to be learned from good practice at other thematically focused repositories such as Timescapes and from national data centres.

All the work of CAVA is documented and available to other projects and services through the CAVA website.

8. Implications

CAVA modelled the development of an audio-visual repository from scratch, including elements of file standardisation, discipline-specific metadata, rights and access management, and sustainability issues. The work of the project is in the public domain and will help to support similar undertakings.

CAVA placed an unusual emphasis, for a HE repository start-up project, on issues of consent and re-use. There is scope for a new body of work to support consensus and uniformity in the management of sensitive content: clarification of legal obligations, collation of best practice, creation of guidance, and current and future conventions for the recording of machine-readable expressions of policy for restricted material could be an interesting set of topics for further exploration.

For the ongoing business case for any repository, it is important to collect evidence of use. CAVA data is primary, and access to it must be restricted, and therefore broad-brush download statistics only tell a small part of the story. CAVA users are now asked to report details of any secondary publications drawing on CAVA data to the project team, for listing on the project website, and to ensure that the repository is acknowledged in those publications. Researchers are also encouraged to include citations to CAVA data sets in publications references lists, where appropriate. Time will tell whether these strategies are successful. However, there seem to be technical opportunities for interactions between primary data and published research outputs, and for the citation of datasets using persistent identifiers. Work in these areas is at an early stage, although progressing quickly¹.

Working in partnership with a national data centre, and in close liaison with a repository facing similar challenges (Timescapes), was fruitful, both in helping to shape the project and in the development of a new model for the digital preservation of HE repository material. Similar partnerships are to be recommended to other UKHE projects working with primary data.

9. References

CAVA Repository: <http://www.ucl.ac.uk/library/cava>

Project website: <http://www.ucl.ac.uk/ls/cava>

CAVA YouTube channel: <http://www.youtube.com/user/CAVArepository>

Guidance

- Format guidance: <http://www.ucl.ac.uk/ls/cava/docs/0709project-report.shtml>
- Troubleshooting: <http://www.ucl.ac.uk/ls/cava/troubleshoot.shtml>
- Metadata guide. (Word). <http://www.ucl.ac.uk/ls/cava/docs.shtml>
- Metadata form. (Excel). <http://www.ucl.ac.uk/ls/cava/docs.shtml>
- FAQ: <http://www.ucl.ac.uk/ls/cava/faq.shtml>
- Citing CAVA: <http://www.ucl.ac.uk/ls/cava/docs/acknowledgement.shtml>
- Guidance to UCL staff for including CAVA costs in funding applications: <http://www.ucl.ac.uk/ls/cava/docs/grant-applications.shtml>
- Standard consent form, adult subjects: <http://www.ucl.ac.uk/ls/cava/docs/consent-form-adult.doc>
- Standard consent form, child subjects: <http://www.ucl.ac.uk/ls/cava/docs/consent-form-child.doc>

Publications and dissemination: <http://www.ucl.ac.uk/ls/cava/dissem.php>

Timescapes project: <http://www.timescapes.leeds.ac.uk/>

¹ For a recent (at the time of writing) overview of issues and initiatives in research data management, see D-Lib Magazine, January-February 2011: <http://dx.doi.org/doi:10.1045/january2011-contents>