IJDC | *General Article*

Archiving Large-Scale Legacy Multimedia Research Data: A Case Study

Claudia Yogeswaran
University College London

Kearsy Cormier
University College London

Abstract

In this paper we provide a case study of the creation of the DCAL Research Data Archive at University College London. In doing so, we assess the various challenges associated with archiving large-scale legacy multimedia research data, given the lack of literature on archiving such datasets. We address issues such as the anonymisation of video research data, the ethical challenges of managing legacy data and historic consent, ownership considerations, the handling of large-size multimedia data, as well as the complexity of multi-project data from a number of researchers and legacy data from eleven years of research.

Received 23 February 2017 ~ *Revision received* 17 August 2017 ~ *Accepted* 17 August 2017

Correspondence should be addressed to Kearsy Cormier, Deafness Cognition and Language Research Centre, University College London, 49 Gordon Square, London, WC1H 0PD, UK. Email: k.cormier@ucl.ac.uk

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

This case study targets researchers and video data archivists working with large, complex, legacy, and/or multimedia research data, and seeks to guide the structured process of its collection, management and preservation for future research purposes. We provide a case study of archival work undertaken at the UCL Deafness, Cognition, and Language (DCAL) research centre to archive research data within the field of deafness and sign language studies. This includes an overview of the planning and implementation process, and considers key issues, resources, deliverables, and problems. Based on an evaluation of the project's successes and failures, we provide a list of recommendations, as well as some input on how these can be implemented in the future to support the archiving of other large multimedia research datasets.

Context

Data archiving has many positive practical and economic implications in the research environment, such as facilitating access to reusable data, aiding knowledge gathering and distribution, and expanding novel research. Archiving can also include a process of anonymisation, adding value to data and research (Korkiakangas, 2014). However, such positive outcomes are difficult to come by, particularly when one must prepare and organise the data prior to archiving.

Legacy Data Collation

Archiving legacy research data presents the initial problem of identifying and detailing information on all data which needs to be gathered, as well as collating the data itself. This can involve liaison work to ascertain the scope of various projects and researchers who worked on them, as well as identifying the location of data; the workload can vary depending on preceding institutional research data management practices and resources.

There are various problems associated with collating legacy research data. Firstly, researchers and staff may have left an institution, taking data with them, and/or have stored information in different locations both on- and off-site. Also, former staff in possession of research data, or knowledge of its location, may no longer be contactable. Secondly, storage upgrades and transfers may have occurred, meaning that data – sometimes incomplete or obsolete – may be replicated across different physical and digital technologies. Finally, in some cases, data may be lost.

A data archiving plan and resources to implement it, relative to the projected timeframe and scale of the project, can ensure these issues are addressed.

Anonymisation of Video Data

Video recorded data is sensitive in nature, and though archiving protocol can take anonymisation (or de-identification) and confidentiality into account and images of participants can be blurred, names coded, and other personal identifiers removed – striving to making a person untraceable from the data presented about them (Saunders et al., 2015) – "complete anonymity cannot be guaranteed" (Korkiakangas, 2014).

Additionally, whereas the inclusion of metadata can be a boon for data reuse, accessibility, richness and quality, description, and comparative analysis and contextual (research value) purposes, there is the possibility that the triangulation of certain metadata can deem a participant identifiable. For example, where an 'anonymised' participant has school, location, and disability information available, it may be possible to pinpoint the person's identity, particularly if the location is small and/or isolated, the school specialised, and/or the disability a key marker.

Sign language video data has the added complexity of containing identifiable personal information, where participants' faces are necessarily shown. The use of the face (including facial expressions, looking at where the eyes are gazing and what the mouth is doing) is essential to sign language research and are impossible to anonymise. This is also true of research using audiovisual spoken language, or multimodality or gesture research, which requires analysis of the face (e.g. Robson, 2011; Haw and Hadfield, 2011; Jewitt, 2012; Parry, 2013).

The data archivist working in research environments such as these (where obscuring facial information in video for anonymisation is impossible) must thus determine the value and usefulness of such video data and metadata for future reuse and research, how it should be managed (i.e. how access to it can be provided appropriately), and how personal and potentially identifiable information can be kept secure. Frequent monitoring and revision of data held can ensure the long-term security, management, ethical regulation, and up-to-date anonymisation of as much data as possible.

Ethical Challenges of Legacy Data

Research ethics practice involves the consideration of quality and integrity in all (planning, acquisition, interpreting, storage, dissemination, and disposal) stages, including the protection and security of sensitive and personal information (Wiles et al., 2012). In most research institutions this is presided over by data management/protection and ethics committees, which provide in-house operational methods which (in the UK) adhere to the government-sanctioned Data Protection Act 1998 (DPA, 1998).

In general, research ethics practices are more stringent now than ever before, especially as the collection and use of electronic data becomes more commonplace. Archiving legacy (particularly multimedia) data can be problematic if research was conducted prior to the rigorous application of ethics protocols, as documents such as ethics agreements and participant consent forms may be lost, or the correct consent (for actions such as storage or dissemination) was not collected in the first place.

Retrospective consent can sometimes be sought, though this is not always ideal; participant contact information may not be available, participants may no longer be contactable, or the process can be so time-consuming that it is not practical. Research participants are also granted the right to withdraw their consent, and this also needs to be deliberated when exercising our responsibility to protect study participants. Retrospective consent can be particularly complex in the case of individuals who are not able to consent for themselves. For example, consent for child research is often given by parents or carers, and in some cases 'informed consent' is given by the child where able (WMA, 2013). This can present a complication when archiving legacy child data; by the time retrospective consent is sought, the child may be old enough to consent themselves, if they can be found.

Even with consent, researchers and data archivists have an ethical responsibility to consider the risks of publishing and storing data, whether multimedia-based, or otherwise (Wiles et al., 2012). This can be made difficult if a researcher has left the

institution with whom the data was gathered, if the institution no longer exists, or the researcher can no longer be consulted (i.e. may be deceased). In such circumstances, current institutional policies must be implemented, and the archivist must make a judgement on whether the data (if it is deemed valuable) can still be published, how accessible it will be and how to manage that access, and security implications of its storage.

Ownership, Copyright, and Associated Legal Issues

Questions of ownership in relation to legacy research data raise legal issues and are always problematic, and the participant, researcher, institution, and data repository all hold some degree of responsibility (IPO, 2014; Deegan and Tanner, 2006). It is important to consider whether any research is under patent or subject to intellectual property law, and if peers need to be consulted (where collaborations have been made). Research that is publicly funded or deemed to be "in the public interest" may mean that the public can consult data under the Freedom of Information Act 2000 (FOIA, 2000), and guidelines need to be implemented to detail legal restrictions (e.g. where the DPA 1998 prevents the dissemination of personal or sensitive data to particular audiences).

Where researchers have left an institution in which data was collected, consent forms need to be reviewed for access privileges. For example, some historic consent forms may inform the participant that collected data will not be used outside of a particular institution, which can be problematic when the researcher of that study leaves for another institution; this can be made more complex when institutional policy dictates the reuse of data by the researcher who collected or led on a project, regardless of whether they remain at the institution in future.

Future ownership of historic data also needs to be clearly defined, and arrangements must be made to resolve ambiguities of whether the participant, researcher, research institution, or repository will be responsible for the storage and management of that data (TNA, 2006). Such conflicts and discrepancies in historic consent and ownership will need addressing when archiving legacy data and will often be resolved sensitively on a case-by-case basis and at an institutional level.

Handling Video Files

Large-size video and its storage

There is a lack of literature, including practical guidance, on the storage and ongoing preservation of large-size research video data. This is not to be confused with 'big data' where the entire dataset is very large. Instead, we refer to individual video data files which are large-sized (e.g. 10GB+) and cannot be uploaded to custom research data archives and repositories (e.g. the UK Data Service, UKDS) because of limits to the file size of uploads and the spatial capacity constraints of servers. This of course results in an entire dataset that is also extremely large, to the extent that some archives may reject it outright from the start.

The lack of suitable research repositories is highly problematic in this and similar fields of research that handle large-size video data, and is reflected all the more so in how little literature there is on the subject of managing and archiving it. It can be argued that as research councils increasingly require that research data be archived for reuse and sharing purposes as a condition for research funding (AHRC, 2016; ESRC, 2015), they must take some responsibility in providing or advising on repositories relevant for

archiving different types of research data (including large datasets), both to aid researchers and address this growing problem.

Format

Although the process of compressing data and converting to different formats can make it easier to upload video and other multimedia files, there is a need to standardise such formatting (e.g. Careless, 2006; Schüller, 2009). However, this is not always possible with video data where the high quality of raw files is desirable for data analysis and reuse purposes (e.g. in sign language research), and appropriate storage for conversion and archiving of multiple formats may not be available or affordable. It may also be necessary to try and anticipate what formats will be most useful overall in the long run, and the time and space needed for creating, converting and storing multiple versions of video.

Deterioration and bit rot

Digital information, much like its physical counterpart, is susceptible to deterioration over time, even if it is not often used. Furthermore, machines and software which read data can become obsolete, and each conversion to another format can lead to a minor loss of information and quality, highlighting the "impermanence of digital storage media" and digital compatibility (Hayes, 1998). Digital surrogates are at present the most effective means of preserving digital information long term; at an additional (on-going) cost and more space-consuming, they can backup data and replace anything that gets lost. In addition, the long-term cost of maintenance, sustainability, and upgrades to data and hardware must also be considered.

Processing and Cataloguing

The organisation of data through processing and cataloguing adds value to a dataset by making it more meaningful, and increases an item's searchability and findability (and thus its usefulness) within the archive. When confronted with raw data which lacks adequate description, or which contains naming complexities (which will involve the decoding of existing information stored elsewhere), a process of identification, description, and organisation will make that data understandable (Marshall et al., 2013).

Cataloguing practice in the information sciences refers largely to the analysis and description of an item, as well as the provision of access to it. With regard to archiving research data, this refers to the accurate and consistent labelling of all descriptors, data, and metadata. Descriptive metadata includes basic attributes such as element, title, author/creator, subject, description, date of creation, place of creation, data type, its relationship to other resources, etc., and administrative metadata includes details necessary for a resource's management (Jordan, 2006).

Certainly, the more descriptive the metadata, the better it can "assist users in discovering resources, evaluating resources, and grouping related resources together. An additional and important function that descriptive metadata serves is that it can be tailored to the resource discovery needs of a specific audience" (Jordan, 2006). This processing and cataloguing is generally a straightforward but time-consuming task, and can be difficult without adequate cataloguing software and/or when there is a large amount and variety of unsorted data.

Summary

In general, the curation of large-size and legacy data is a challenge increasingly common within many research programmes and at many research centres. Although there is a growing body of literature (in information sciences and other subject-specific academic fields) guiding researchers and archivists on the curation of research data and detailing such practice (Schubotz et al., 2011; Bardyn et al., 2012; Marshall et al., 2013), there is a notable lack of guidance and literature exploring the difficulties of managing research datasets with large files, and/or that involve video data, and/or that have been created retroactively for archiving legacy data that already exists. The dearth of guidance on these combined issues particular to the DCAL data archiving project made our task a more difficult one.

Case Study: DCAL Research Data Archive

Background

Funded by the Economic and Social Research Council (ESRC) (Phase I January 2006 – December 2010 and Phase II January 2011 – December 2016) and based at University College London, DCAL is the largest deafness and sign language research centre in Europe, bringing together leading Deaf and hearing researchers in the fields of sign linguistics, psychology, and neuroscience (ESRC, 2017a; ESRC, 2017b). During these 11 years, researchers worked on 94 DCAL-funded and DCAL-associated projects, resulting in over 600 publications, and generating over ten terabytes of research data.

The archiving project we report here was undertaken in the context of the funder's requirement that all DCAL research data should be submitted to the UKDS ReShare repository by the time the funding ends. Because the funder's specific requirements for data submission developed and changed throughout the DCAL lifespan, a period of six months was put aside for this project near the end of the centre funding (June – December 2015) to consolidate all centre research data to meet the data submission requirement. A Data Archive and Management Officer (the first author, CY) was employed to undertake the task of collating and preparing data, and for creating a Data Management and Archiving Policy to support future data collection management.

Project Planning and Implementation

When embarking on this project, it was important to identify the objective and expected outcomes, what resources were available, and any issues and risks relating to collating, organising, and cataloguing research data for the archive. This initial assessment allowed us to develop a plan and timeline detailing deliverables to meet the intended outcomes.

Objective and expected outcomes

The aim of the DCAL archiving project was to create a searchable and accessible electronic repository to hold all DCAL research data. This would meet not only the ESRC funding requirement, but would also make usable all research data related to projects emanating from DCAL over the previous 11 years.

The principle expected outcomes and benefits of this project were to: meet the funding requirements of the ESRC; create an organised, searchable, and accessible research data archive which is user-friendly with a front-facing support system for assisted access; impact on data sharing and knowledge distribution and dissemination on a larger, global scale, thus ensuring the advancement of knowledge, advancing learning and education internationally, and supporting researchers in and out of the field at all levels, and; develop a Data Archiving and Management Policy (DAMP) (Cormier and Yogeswaran, 2017), including work flow guidelines, to support future research data collection within DCAL.

Resources

The principle resources required and available for the project were DCAL researchers and staff. Liaising with them ensured that a full list of what needed to be archived could be drawn up, and also supported the development of the DAMP (Cormier and Yogeswaran, 2017). The ESRC, UKDS, and UK Data Archive (UKDA) websites and their staff, were also informative and helpful, giving guidance on archiving and data format requirements. UCL Digital Collections (which already hosted the British Sign Language Corpus via CAVA, a human Communication Audio-Visual Archive for UCL) and the Digital Curation Manager at UCL Library Services offered support in terms of building the archive and helping to plan the development of the repository. The IT Officer at DCAL provided support with data migration (including the building of servers) and video conversion. The UCL Research and Ethics Committee, UCL Legal Services, and UCL Digital Collections' Research Data Support Officer were available for help regarding consent, ethics, and permissions queries. Finally, with regards to the budget, DCAL funds were available for the development of the research data archive.

Deliverables

Specifying the scope for the project (in relation to the objective and resources available) allowed us to determine deliverables for the six-month timeframe. This included: the consolidation of research data by gathering, organising, and preparing datasets with assistance from DCAL directors, researchers, and other staff involved in creating and maintaining the data and from the ESRC, UKDA, and UKDS regarding data formats and information types; clarifying permission, consent, and ethics guidelines in terms of DCAL research; writing a Data Archiving and Management Policy to support future research data management practices at DCAL, and; creating and implementing a DCAL Research Data Archive by liaising with UCL Digital Collections.

Mitigating risks

The main risk we had to consider in the preliminary stages was that, due to the nature of research work, some researchers would be difficult to contact, or unable to respond immediately to requests about data (e.g. location, transfer, and preparation) due to professional workloads, leave, or other extenuating circumstances. It was therefore important to start contacting researchers as early as possible.

There was also the possibility that researchers may have large datasets, or would have difficulty preparing data to particular specifications. Therefore, we resolved to make the preparation process as smooth as possible by providing sufficient support and guidelines, and planning for the Data Archive and Management Officer to take on some

data-preparation support work (e.g. digitising materials, sorting data into file directories, adding metadata, etc.) on an interim basis.

Other risks to consider were: (1) the budget, since although some DCAL funds were available for archiving, we did not initially know what total costs would be needed for the entire archiving project from start to finish, and (2) the digitisation of physical documents, which was highlighted as being potentially time-consuming. The chief ramification for not finding solutions to these risks from the outset was that archiving would either not be fundable or would fall behind schedule and the ESRC funding requirement would not be met.

Key Issues

When the archiving project was underway we encountered some problems relating to the understanding of ESRC funding requirements and data submission guidelines, data collation and organisation (including support provided to researchers), digitisation practices, the archiving of multimedia research data, finding and building the repository, and ethical and legal concerns.

Understanding ESRC data submission guidelines

It was necessary to identify what formats and types of information needed to be archived. The ESRC, UKDS, and UKDA websites contained a considerable amount of information on how to correctly prepare research data, but they tended to be focused on research projects rather than centres, and although comprehensive, were sometimes difficult to navigate. It was also difficult to find out about *where* data needed to be deposited. We quickly learned that it was not possible to upload directly to the UKDS ReShare repository, as encouraged, due to file size limitations (see below). Also, information about alternative suitable repositories, with an emphasis on video data, was not detailed. Liaising with representatives of various organisations/archives was therefore very useful for getting clarification on the types of data and detail of metadata required, and alternative options for depositing data.

Data collation

To begin the process of consistent data collation across DCAL, the first task was to identify key projects and associated Principal Investigators (PIs) and researchers. One problem was that many DCAL projects were related to each other and the relationships were not always immediately obvious. For identifying the scope of individual projects, we relied heavily on project titles. However, project titles often shift at various stages of research, particularly as projects evolve – sometimes splitting into new projects – and discoveries are incorporated thereby changing the direction of a project. By meeting with PIs it was possible to establish a full and definitive list of project titles, as well as a record of all researchers associated with each onto a spreadsheet.

We liaised with PIs and research staff to determine what project data and metadata they held, and where it was stored, creating an inventory in the process. This was relatively straightforward with staff still employed at DCAL, but where researchers had left, it was necessary to recover contact details which was not always easy.

Occasionally, physical and analogue data had been stored in different offices, cupboards, and filing cabinets around the building and other locations. An inventory allowed us to reorganise, document, and safely and securely store data from finished projects in one physical location.

Afterwards, a digitisation plan was established consisting of two parts: scanning paper data to reduce physical storage space within the research centre, and converting video data from analogue media to digital form (this latter point is discussed further below). Digitising data allowed us to migrate and store all information on a dedicated directory on the DCAL server.

Following liaison with UCL Information Governance, a SFTP server was set up by the DCAL IT Officer, and this ensured that institutional security protocols were met. The SFTP server allowed us to receive (often large) data transfers from staff off-site, so our full research data collection from dispersed locations, devices, and media was consolidated on the central server within DCAL and ready for organisation and preparation.

Data preparation

As data began to be collated in digital format, it became evident that not everything was clearly or consistently organised or labelled across researchers/projects, and although they were organised in a meaningful way for the immediate researchers on that project, they were often not comprehensible to researchers outside of that project. Digitised documents and data required labelling anew. A thorough regiment needed to be implemented to ensure that the correct data was collected, consistently organised and prepared, systematically catalogued, and necessary metadata recorded and inputted. The ESRC had certain requirements on how data and metadata had to be prepared, but sometimes these requirements were not clear and thus needed to be clarified.

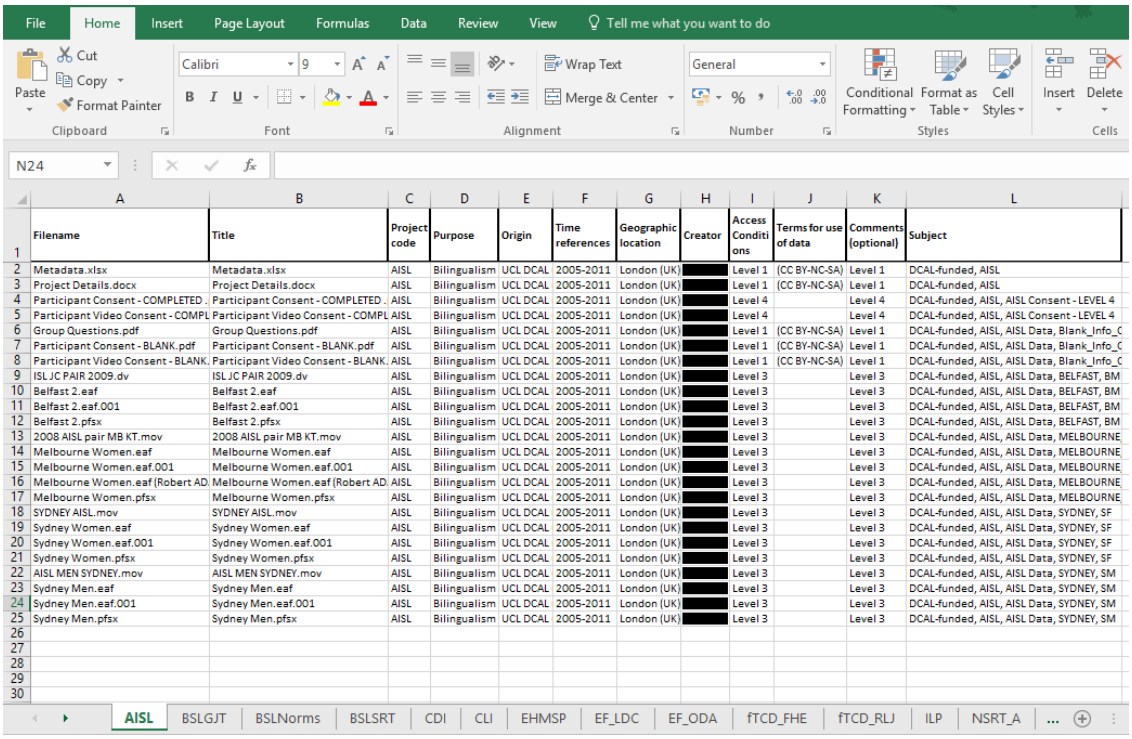
A guidance document (see Cormier and Yogeswaran, 2017) was drawn up and distributed to researchers so that they could prepare their data for archiving, with the intention of making the process as easy as possible for them. This document requested from each researcher: the project title; an identifying project code; a short description of the project for the archive; a long description of the project dataset for the archive; the research theme or strand it fell under; who the PI and associated researchers and authors were; information about the location of research data and whether any data needed to be digitised; copies of project information sheets and blank consent forms; identification of applicable data restrictions and terms of use, and; a list of all projects each researcher or PI was involved on. Compiling this information allowed us to organise all project (and collection-level) information.

Tailoring descriptive metadata to academics and professionals working in DCAL's various research fields would require bespoke and specialised data entry (Jordan, 2006). Given the incredible volume of data and time limitations, however, it was decided that metadata would remain at a minimum, following the IMDI (ISLE Meta Data Initiative) metadata standard (with field names detailed and explained in Figure 1 and also in Cormier and Yogeswaran, 2017), which describes multimedia and multimodal language resources (TLA, 2010).

Metadata for DCAL (video and non-video) data (IMDI standard).			
#	Field name		Field description
	Part A (one column for each file)		
A01	✓	Purpose	Description of why the data was carried out/project title
A02	✓	Origin	Where the data comes from
A03	✓	Time.References	When the data was created
A04	✓	Geographic.Location	Where the data was compiled (i.e. London, UK)
A05	✓	Creator	List all authors of the data (in citation order)
A06	✓	Access.Conditions	Restriction level (see DCAL restriction levels)
A07	✓	Terms	Any terms of use
A08		Comments	Optional

Figure 1. Project metadata (IMDI).

For accessibility and searchability purposes, DCMI (Dublin Core Metadata Initiative) metadata terms were added, and included file name, file title, a unique code for each project, terms for use of data, and level of access restriction (see Figure 2). The Data Archive and Management Officer was on-hand to help with this, and was able to undertake some of the work where researchers had explained the datasets.



Filename	Title	Project code	Purpose	Origin	Time references	Geographic location	Creator	Access Conditions	Terms for use of data	Comments (optional)	Subject
Metadata.xlsx	Metadata.xlsx	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 1	(CC BY-NC-SA)	Level 1	DCAL-funded, AISL
Project Details.docx	Project Details.docx	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 1	(CC BY-NC-SA)	Level 1	DCAL-funded, AISL
Participant Consent - COMPLETED .	Participant Consent - COMPLETED .	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 4		Level 4	DCAL-funded, AISL, AISL Consent - LEVEL 4
Participant Video Consent - COMPL	Participant Video Consent - COMPL	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 4		Level 4	DCAL-funded, AISL, AISL Consent - LEVEL 4
Group Questions.pdf	Group Questions.pdf	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 1	(CC BY-NC-SA)	Level 1	DCAL-funded, AISL, AISL Data, Blank_Info_C
Participant Consent - BLANK.pdf	Participant Consent - BLANK.pdf	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 1	(CC BY-NC-SA)	Level 1	DCAL-funded, AISL, AISL Data, Blank_Info_C
Participant Video Consent - BLANK.	Participant Video Consent - BLANK.	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 1	(CC BY-NC-SA)	Level 1	DCAL-funded, AISL, AISL Data, Blank_Info_C
ISL JC PAIR 2009.dv	ISL JC PAIR 2009.dv	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, BELFAST, BM
Belfast 2.eaf	Belfast 2.eaf	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, BELFAST, BM
Belfast 2.eaf.001	Belfast 2.eaf.001	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, BELFAST, BM
Belfast 2.pfsx	Belfast 2.pfsx	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, BELFAST, BM
2008 AISL pair MB KT.mov	2008 AISL pair MB KT.mov	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, MELBOURNE
Melbourne Women.eaf	Melbourne Women.eaf	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, MELBOURNE
Melbourne Women.eaf.001	Melbourne Women.eaf.001	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, MELBOURNE
Melbourne Women.eaf (Robert AD.	Melbourne Women.eaf (Robert AD.	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, MELBOURNE
Melbourne Women.pfsx	Melbourne Women.pfsx	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, MELBOURNE
SYDNEY AISL.mov	SYDNEY AISL.mov	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SF
Sydney Women.eaf	Sydney Women.eaf	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SF
Sydney Women.eaf.001	Sydney Women.eaf.001	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SF
Sydney Women.pfsx	Sydney Women.pfsx	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SF
AISL MEN SYDNEY.mov	AISL MEN SYDNEY.mov	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SM
Sydney Men.eaf	Sydney Men.eaf	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SM
Sydney Men.eaf.001	Sydney Men.eaf.001	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SM
Sydney Men.pfsx	Sydney Men.pfsx	AISL	Bilingualism	UCL DCAL	2005-2011	London (UK)	[redacted]	Level 3		Level 3	DCAL-funded, AISL, AISL Data, SYDNEY, SM

Figure 2. Project metadata (DCMI).

A file directory hierarchy was constructed to support the data organisation (shown in Figure 3). Broadly speaking, data were organised by whether they were linked to projects that were DCAL-funded (i.e. research directly funded by the DCAL centre grant) or DCAL-associated (i.e. research undertaken at the research centre by staff, but not directly funded by the DCAL centre grant). DCAL-funded research was prioritised and it was decided that DCAL-associated research would also be archived once the ESRC requirements for DCAL were met. With each category, the research strand/theme

was distinguished, and within these the project titles (marked by a project code). At the project level, research data and the affiliated documentation within directories was labelled consistently (finalised structure shown in Figure 4). This also maintained a clear structure when uploading files to the archive.

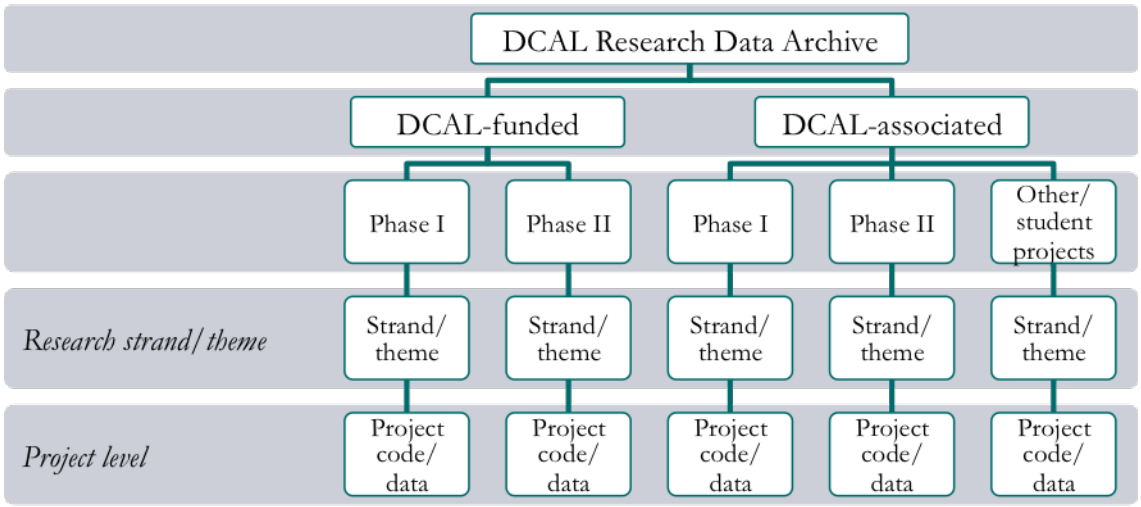


Figure 3. File directory hierarchy for project data organisation.

```
[EthicsPI_EthicsNumber + 'Programme' Folder] Only PI and co-applicants have full access.
PI/co-applicants may grant access to other individuals for specific projects.
     |---- [Ethical approval, signed/scanned risk assessments etc. at research programme
     level]
      [Project_Name + 'Project' Folder] Only group-project members have access
        |---- [… more experiments…]
        |---- [documentation applicable at a project level] Project Metadata form for
        archiving
        |---- [Experiment_Name + 'Exp' Folder]
              |---- [Participants List] -Optional if using DCAL Participant Database
              Sensitive Data
              |---- [Consent Forms] Sensitive Data
              |---- [Questionnaires] Not sensitive data - Use only SubjectIDs
              |---- [Participant Metadata] Not sensitive data - Use only SubjectIDs
              |---- [All Info] All the Docs to understand, run & perform the exp.:
                    |---- blank info sheets, consent forms & questionnaires
                    |---- briefing materials/emails/adverts/instructions
                    |---- design layout of exp.
              |---- [Stim] Source materials for the exp. like pics, videos, etc.
              |---- [Output] Raw data collected: logs, videos, etc.
              |---- [Final Output] Processed data as was used in publication
              |---- Outside folder: Program(s) code or scripts used in exp.
```

Figure 4. Project directory hierarchy.

Consent, permission, and ethics

Although it was clear from the outset that DCAL research data would be required to be deposited at the end of the centre grant, the specific ESRC requirements about depositing were not clear at the start; even if they had been, they changed over the course of the life of DCAL. Thus, consent forms used for some projects (particularly the earlier projects) had not mentioned data retention and data sharing. Ethics concerns

regarding future data storage were also highlighted, and where researchers had not obtained explicit consent to keep research data for future reuse, and where we could not seek retrospective consent due to time restrictions, we had to liaise with UCL Legal Services and UCL Digital Collections' Research Data Support Officer who advised that non-anonymisable data could be archived if access was restricted to project PIs only.

The archive also implements a 'membership' model of access, where data that is not publicly available can only be viewed by project PIs or researchers. Other research data can be made available to 'approved' researchers, who must contact the project PI in order to gain the appropriate permissions, make a contractual agreement to ensure correct and ethical use, and upon meeting requirements, will then be granted access to the relevant dataset within the archive for a certain timeframe. This ensures that personal information about participants remains secure and enables continued research.

Furthermore, the UCL Research Data Policy (UCL, 2013), UCL Records Retention Schedule (UCL, 2015a), UCL Staff IPR Policy (UCL, 2015b), and DPA 1998 all contained guidance on how to store, manage, and review research data in the long-term and maintain its integrity. Such a solution was agreeable in meeting our funding requirement and for the support of future research. For example, in the past where researchers have been unable to view research data due to restrictions, they have been able to contact PIs who can describe or show a modelled/anonymised version of that data in a way which protects the identity of participants and still deems that research data reusable and useful.

Data loss and security

To mitigate the risk of a loss of data, the DCAL IT Officer created a digital surrogate of data during the archiving process, to replace anything that might get lost. UCL Digital Collections, similarly, maintains data through the institutional Information Data Safe Haven (IDHS), which has been certified to the ISO 27001 information security standard. Annual reviews of the DCAL archive will include security protocols for research data stored with UCL Library Services. Alongside their electronic counterparts, some original paper documents, such as consent forms, have also been retained and stored securely.

Video research data

Sign language research necessarily involves video data. Most research video files handled at DCAL are very large, ranging from 100MB to 10GB each. In-house hardware is able to support the storage of this, but as it was necessary to archive our data for future reuse, finding a repository which accepted such large files proved to be very difficult. In some cases, video files were compressed to reduce file size, but even following this process, some files were still very large, and converting to other (sometimes lossy) formats was not desirable as high quality video is required for data analysis.

Research data formats

Data collected over the life of DCAL was stored in various formats, both physical and digital. Papers and objects and their documentation were locked securely within the research centre and off-site – this included analogue media such as Video8 videocassettes and VHS cassettes, and digital media such as MiniDV tapes, Zip, CD, DVD, floppy disks (various sizes) and HDD, all with various audio and video codecs. Other formats for research files included those which are readable only by specific software (e.g. SPSS, MATLAB, etc.). Digital file formats and data file extensions were

multifarious and wide-ranging, and it was important to consider consistency across computer-readable formats (Hughes, 2003).

Finding a suitable repository

There was little infrastructure available to support the archiving of research data – particularly multimedia research data – in major UK research archives, and there were not many research depositories wishing to take on or store the quantity and large size of video data that we had and assign Digital Object Identifiers (DOIs) which make data permanently citable (Corti, 2012). The UKDA's UK Data Service (which includes the former Economic and Social Data Service) is the repository recommended by the ESRC, but would only accept data deposits consisting of files with a size of maximum 2GB each, whereas many of our video files were over 10GB per file.

We also consulted institutional research data archives such as the Oxford University Research Archive (ORA) and archives at the University of Cambridge and University of Bristol, amongst others. Although they have good support frameworks and guidelines for archiving funded research data, they only accept data collected at those respective organisations.

The DCAL-associated BSL Corpus Project had previously deposited with UCL Digital Collections, and this was put forward as a possible archival solution. UCL Digital Collections accepts large-size video files and assigns DOIs to datasets, and choosing this home option also had the added benefit of in-house research management, ethics, and legal support. Following this rationale and after some initial consultations with the Digital Curation Manager, it was decided that this would be the repository to archive with.

Building and implementing the archive

Working with the UCL Digital Collections, we were given guidance on depositing requirements such as file labels, costs (for building and maintaining research data on UCL Library Services' servers for the long-term), migration protocols for securely moving data across servers, and metadata and file directory layouts. We also needed to provide information on the layout structure of projects, as well as project descriptions to complement the archive and its searchability once all the data was live. Archiving with UCL Library Services ensured adherence to the ISO 14721 information security standard.

Since 2016, the DCAL Research Data Archive has been listed on the ReShare repository (DCAL, 2016a) and is available through UCL Library Services' Digital Collections (DCAL, 2016b) as shown in Figures 5 and 6. Although at the time of writing, UCL Digital Collections are still working to resolve the access and permission systems, the archive was submitted on time to the ESRC and now holds the fully catalogued collection of DCAL-funded research data from 2006-2016.

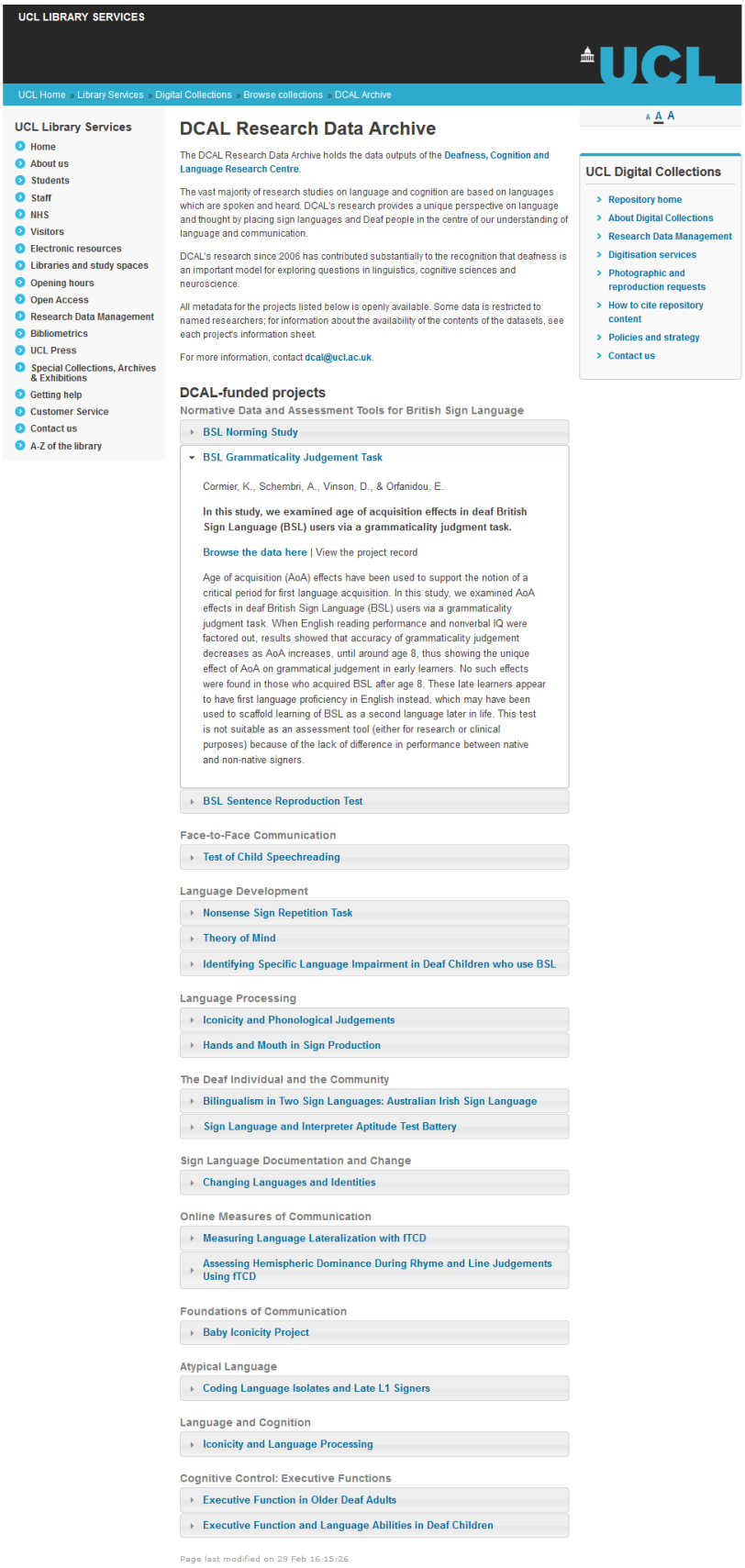


Figure 5. Screenshot of DCAL Research Data Archive (DCAL, 2016b).

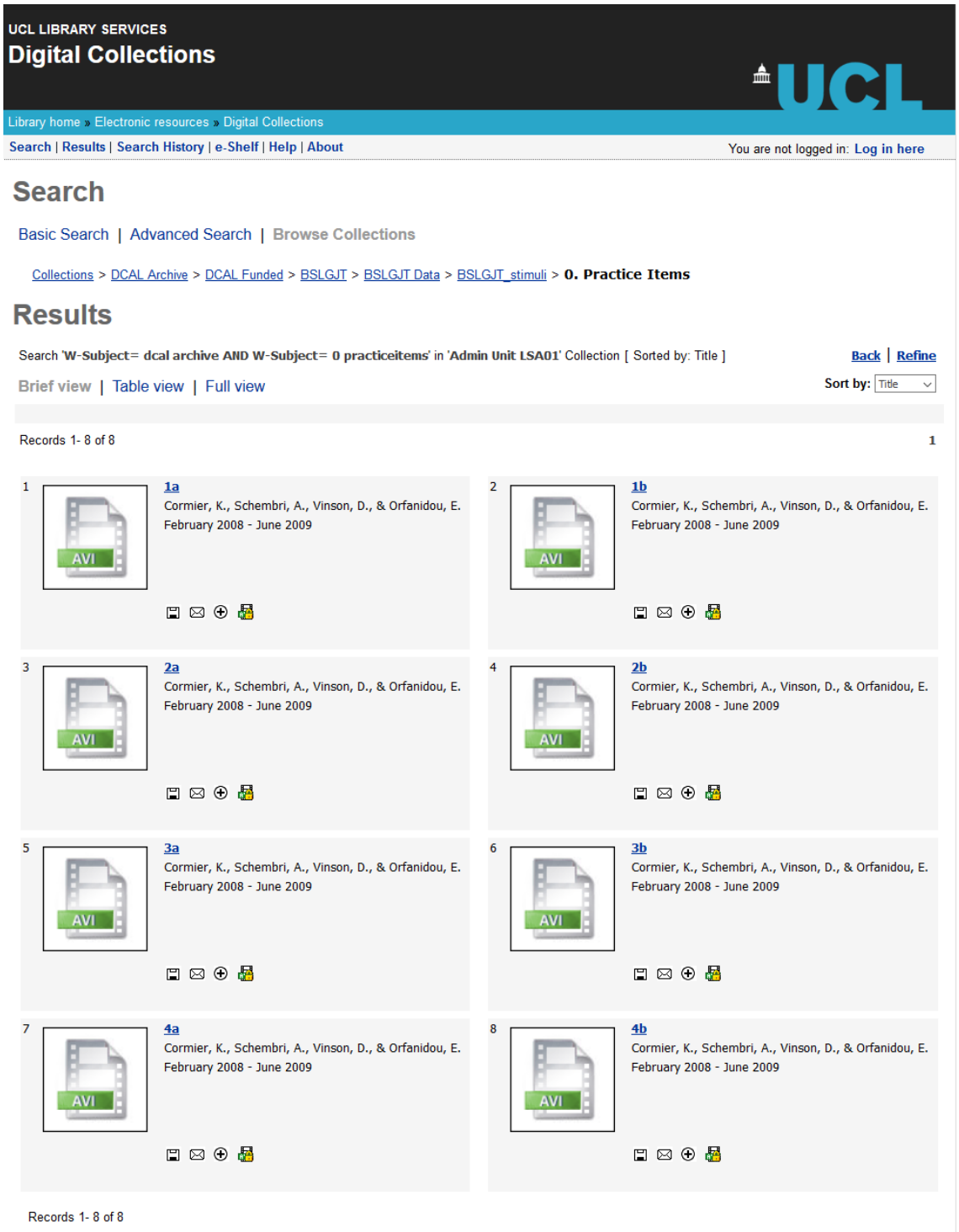


Figure 6. Screenshot of DCAL Research Data Archive item examples (DCAL, 2016b).

Evaluation

The project was successful in meeting all the objectives defined at the outset, and has brought some key points to light.

There were some difficult aspects, such as the multi-pronged rather than linear process, with complex, interrelated projects, many with overlapping timelines and

resource factors, to consider. However, having a strong plan in place from the outset of the project was beneficial, as was the support and guidance offered to us by UCL staff and policies, which made this undertaking much easier.

By the end of the project, points of importance we identified relate to database and archive building – particularly the value of metadata and searchability – in research data management, and the need to understand the interdisciplinary nature of both research and archiving practice, where people with different expertise increasingly need to work together (e.g. the information professional's role to support researchers and knowledge gathering/dissemination).

Another salient detail is the growing need to understand video and multimedia data in the research environment, including the intricacies involved in collecting, converting, storing, archiving, and reusing it, especially when it is not anonymisable. Research ethics and legal concerns also need consideration.

Future Data Management at DCAL

Following recommendations and findings from the case study, DCAL has put into place a Data Archive and Management Policy (Cormier and Yogeswaran, 2017) which supports future Data Managers at DCAL with research data management, and includes guidelines for helping research staff with preparing data for archiving. The document highlights the need for regular reviews of various related government and institutional policies, and a closer collaboration of information and research professionals at the centre to ensure an efficient and timely archiving process, including instructions on how to effectively archive new and existing (e.g. previous DCAL-associated) projects.

Recommendations

This case study provides a number of recommendations which will aid researchers and information professionals working with multi-project and/or multimedia and/or legacy research data.

- **Role of funders:** As the need to archive research data is more commonplace in funding requirements, funders must give better guidance on archiving practice and repositories for different types of research data (including large, complex datasets). As requirements change, research councils should work with their longer-term investments to ensure that changes are implemented efficiently and appropriately.
- **The importance of a Data Manager from the start:** Research involving multiple projects needs oversight by a Data Manager who regularly reviews data management practices, including consideration of ethical, copyright, and legal issues. This should be considered part of the cost of archiving.
- **Guidance for researchers:** The Data Manager should provide guidance on how to collect, prepare, and archive research data – e.g. via a Data Archive and Management Policy (DAMP) (Cormier and Yogeswaran, 2017).
- **Currency of the DAMP:** The Data Manager should keep the DAMP up-to-date with regular reviews of institutional research data policies (e.g. UCL, 2013), records retention schedules (e.g. UCL, 2015a), staff IPR policies (e.g. UCL,

2015b), DPA 1998, and other protocols followed by institutional information governance services to ensure researchers have the relevant information to proceed with data collection and management.

- **Privacy, data access, and data stewardship information for PIs and researchers:** Induction and annual appraisals for research staff should cover information governance such that they receive up-to-date data risk management to help ensure buy-in of the DAMP.
- **Importance of metadata:** Given that metadata influences the authenticity and integrity of a data collection, as well as its usefulness, the DAMP should include some standard metadata fields to be used by all projects within the organisation even if there are additional/optional metadata fields that are project-specific. Metadata guidance should also include information relating to funder requirements.
- **Dealing with legacy metadata:** Researchers should be advised on how historical descriptive metadata can be processed for future use ('future-proofing'), including what is good/bad/adequate catalogue data, and whether 'bad' metadata can limit the practical and ethical commitments of collections in the longer term.
- **Technical issues with video:** When archiving video and multimedia research data, adequate funds must be put aside for storage of large-size data, and a video specialist should be employed.
- **Consent issues with multimedia data:** Anonymisation of multimedia research data (via video, images, audio) is emerging as a problematic area, and needs consideration and guidance for possible reuse. A wide range of possible future uses should be considered when seeking consent for collection of such data in order to reduce the need to go back to get consent for later use retrospectively (which is sometimes difficult or impossible). This can be addressed in consent forms with wording relating to retention and reuse.
- **Archiving legacy data:** Using a flexible research repository that allows different levels of permissions for different subsets of data is one way of ensuring that legacy data can be archived to meet funder requirements and, where possible, be reused.

Conclusion

This case study has outlined the creation of the DCAL Research Data Archive at University College London and the range of challenges associated with archiving large-scale legacy multimedia research data. These include the anonymisation of video research data, the ethical challenges of managing legacy data and historic consent, ownership considerations, the handling of large-size multimedia data, as well as the complexity of multi-project data from a number of researchers and legacy data from eleven years of research. These challenges can be mitigated with planning by research centres from the start through investment in a Data Manager but it is still possible to archive legacy data even if data management practices were variable during the life of the research centre.

Acknowledgements

We thank UCL Library Services, UCL Research and Ethics Committee and UCL Legal Services for their support in developing the DCAL Research Data Archive. We particularly thank the DCAL IT Officer Dani Diaz and the UCL Digital Curation Manager Matt Mahon for their assistance and guidance. This work was supported by the Economic and Social Research Council of Great Britain (Grant RES-620-28-0002, Deafness, Cognition, and Language Research Centre (DCAL)).

References

Arts and Humanities Research Council (AHRC). (2016). *Research funding guide: Version 3.6*. Swindon, UK: Arts and Humanities Research Council. Retrieved from http://www.ahrc.ac.uk/documents/guides/research-funding-guide/

Bardyn, T.P., Resnick, T., & Camina, S.K. (2012). Translational researchers' perceptions of data management practices and data curation needs: Findings from a focus group in an academic health sciences library. *Journal of Web Librarianship, 6*(4), 274-287. doi:10.1080/19322909.2012.730375

Careless, J. (2006). PBS, Library of Congress work to preserve public television. *Government Video, 17*(9), 18.

Cormier, K. & Yogeswaran, C. (2017). *DCAL Data Archive and Management Policy (DAMP)*. London, UK: Deafness, Cognition and Language (DCAL) Research Centre, University College London. Retrieved from http://discovery.ucl.ac.uk/id/eprint/1559576

Corti, L. (2012). Recent developments in archiving social research. *International Journal of Social Research Methodology, 15*(4), 281-290. doi:10.1080/13645579.2012.688310

Data Protection Act (DPA). (1998). *Data Protection Act 1998*. Retrieved from http://www.legislation.gov.uk/ukpga/1998/29

Deafness, Cognition, and Language (DCAL) Research Centre. (2016a). *UK Data Service ReShare: DCAL Research Data Archive* [Data set]. Colchester, UK: UK Data Service. doi:10.5255/UKDA-SN-852138

Deafness, Cognition, and Language (DCAL) Research Centre. (2016b). *UCL Library Services: DCAL Research Data Archive*. doi:10.14324/000.ds.DCAL

Deegan, M. & Tanner, S. (2006). *Digital preservation*. London, UK: Facet Publishing.

Economic and Social Research Council (ESRC). (2015). *ESRC research data policy*. Swindon, UK: Economic and Social Research Council. Retrieved from http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy

Economic and Social Research Council (ESRC). (2017a). *Deafness, Cognition and Language Research Centre (DCAL): Phase I, 2006 – 2010*. Retrieved from http://www.researchcatalogue.esrc.ac.uk/grants/RES-620-28-6001/read

Economic and Social Research Council (ESRC). (2017b). *Deafness, Cognition and Language Research Centre (DCAL): Phase II, 2011 – 2016*. Retrieved from http://www.researchcatalogue.esrc.ac.uk/grants/RES-620-28-0002/read

El Emam, K., Rodgers, S., & Malin, B. (2015). Anonymising and sharing individual patient data. *BMJ: British Medical Journal*, 350, 1-6. doi:10.1136/bmj.h1139

Freedom of Information Act (FOIA). (2000). *Freedom of Information Act 2000*. Retrieved from http://www.legislation.gov.uk/ukpga/2000/36

Haw, K., & Hadfield, M. (2011). *Video in social science research: Forms and functions*. Oxford, UK: Routledge.

Hayes, B. (1998). Bit rot. *American Scientist, 86*(5), 410-415. Retrieved from http://www.jstor.org/stable/27857092

Hughes, L. (2003). *Digitizing Collections: Strategic issues for the information manager*. London, UK: Facet Publishing.

Intellectual Property Office (IPO). (2014). *Ownership of copyright works*. Retrieved from https://www.gov.uk/guidance/ownership-of-copyright-works

Jewitt, C. (2012). *An introduction to using video for research. National Centre for Research Methods Working Paper, 03/12*. Swindon, UK: Economic and Social Research Council. Retrieved from http://eprints.ncrm.ac.uk/2259/4/NCRM_workingpaper_0312.pdf

Jordan, M. (2006). *Putting content online: A practical guide for libraries*. Oxford, UK: Chandos Publishing.

Korkiakangas, T. (2014). Challenges in archiving and sharing video data: Considering moral, pragmatic, and substantial arguments. *Journal of Research Practice*, *10*(1), 1-18. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/454/350

Marshall, B., O'Bryan, K., Qin, N., & Vernon, R. (2013). Organizing, contextualizing, and storing legacy research data: A case study of data management for librarians. *Issues in Science and Technology Librarianship*, 74. doi:10.5062/F4K07270

Parry, R. (2013). Video-based conversation analysis. In I. Bourgeault, R. Dingwall, & R. de Vries (Eds.), *The SAGE Handbook of Qualitative Methods in Health Research*. London, UK: Sage.

Robson, S. (2011). Producing and using video data in the early years: Ethical questions and practical consequences in research with young children. *Children & Society*, *25*(3), 179-189. doi:10.1111/j.1099-0860.2009.00267.x

Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: Challenges and compromise in practice. *Qualitative Research, 15*(5), 616-632. doi:10.1177/1468794114550439

Schubotz, D., Melaugh, M., & McLoughlin, P. (2011). Archiving qualitative data in the context of a society coming out of conflict: Some lessons from Northern Ireland. *Forum: Qualitative Social Research, 12*(3), 1-19. Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs1103133

Schüller, D. (2009). Video archiving and the dilemma of data compression. *International Preservation News*, 47, 5-7. Retrieved from https://www.ifla.org/files/assets/pac/ipn/47-may-2009.pdf

The Language Archive (TLA). (2010). *IMDI metadata*. Retrieved from https://tla.mpi.nl/imdi-metadata

The National Archives (TNA). (2006). *Loan (deposit) agreements for privately owned archives*. Surrey, UK: The National Archives. Retrieved from https://www.nationalarchives.gov.uk/documents/archives/loanagreement.pdf

University College London (UCL). (2013). *UCL research data policy*. London, UK: University College London. Retrieved from http://www.ucl.ac.uk/isd/services/research-it/documents/uclresearchdatapolicy.pdf

University College London (UCL). (2015a). *UCL records retention schedule*. London, UK: University College London. Retrieved from http://www.ucl.ac.uk/library/docs/retention-schedule.pdf

University College London (UCL). (2015b). *UCL staff IPR policy*. London, UK: University College London. Retrieved from https://www.ucl.ac.uk/library/copyright/ipr

Wiles, R., Coffey, A., Robinson, J., & Heath, S. (2012). Anonymisation and visual images: issues of respect, 'voice' and protection. *International Journal of Social Research Methodology, 15*(1), 41-53. doi:10.1080/13645579.2011.564423

World Medical Association (WMA). (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, *310*(20), 2191-4. doi:10.1001/jama.2013.281053