DCC SCARP Project Case Study on Roles and Reusability of Video Data in Social Science

Workshop Report– Building and Curating Online Video Data Corpora

April 22, 9.45am - 2pm
Digital Curation Centre, room 7.02 Appleton Tower

Aims of the workshop were to:

- Raise mutual awareness of research communities' practices and needs for archiving, sharing and re-using digital video data.
- Identify how local and national research data services may contribute to the infrastructure for video data curation.

Participation was by invitation to -

- Academics/researchers using video as data in observational, experimental or practice-based research that employs social research methodologies.
- Institutional providers of relevant research data services and tools.

There were 19 people taking part –

- Peter Burnhill, Director EDINA
- Cuna Ekmekcioglu, Research Computing
- Mark Hartswood, Social Informatics Group
- Sarah Higgins, Standards Advisor, Digital Curation Centre
- Robin Hill, Visual Cognition Group, Psychology
- Nicola Hillhouse, Vidiowiki
- Jane Jacobs, Chair in Cultural Geography, School of Geosciences
- Eric Laurier, Senior Research Fellow, Human Geography Research Group
- John Lee, Senior Lecturer, School of Architecture, Culture and Environment
- Stuart Macdonald, Assistant Data Librarian, EDINA
- Abdul Majothi, Academic Liaison Director, College of Humanities and Social Science
- Andy Pryde, Information Services Multimedia Team
- Susen Rabold, Human Communication Research Centre
- Robin Rice, Data Librarian, EDINA
- Jean Ritchie, Director Edinburgh Compute and Data Facility
- Chris Rusbridge, Director, Digital Curation Centre
- Eduardo Serafin, e-Learning Officer, School of Geosciences
- Ignaz Strebel, Research Fellow, Human Geography Research Group
- Angus Whyte, Research Officer, Digital Curation Centre

Apologies were received from Morag Watson, Digital Library and James Stewart, ISSTI.

Programme

Introduction – aims of the workshop

Chris Rusbridge introduced the agenda and the aims of the SCARP Project to engage with researchers to better understand their curation practices. Angus Whyte outlined the background to the case study on 'roles and reusability of video data'. Potential opportunities and challenges for curation were suggested by informal data sharing and re-use of 'corpora' of examples in fields investigating human interaction, where there was some apparent convergence in projects and tools aiming to enable web-accessible archiving. The study had initially involved Eric Laurier and had 'snowballed' to include others associated with the 'SEDIT' group convened by him, along with respondents to a survey of the College of Social and Political Sciences.

Current uses of digital video data and the case for online corpora

Researchers gave short 'scene setting' presentations, outlining how they are using video as observational or experimental data, how their research traditions or topics influence the technologies they use, and the benefits or constraints on making archived video (and related data) accessible online.

Eric Laurier *- Human Geography, SEDIT (Scottish Ethnomethodology Discourse, Interaction and Talk) ' Where the Video Goes'*

Eric identified his work, which routinely uses camcorders to gather video and employs digital video clips as a main data source, with French linguist Lorenza Mondada's video studies of spatial aspects of interaction in various settings including surgical operating theatres. He has taken a similar approach in recent ESRC funded projects including 'Habitable Cars' and (currently) 'Assembling the Line'. The former investigated interaction in cars, the latter in the work of professional and amateur video editors. In these projects around 10% of recorded footage is selected and 'logged', and this further reduced as clips are analysed in 'data sessions' with colleagues, and a selection worked up for publication.

The corpus resulting from the cars project is being reused in other researcher's work, and was submitted to the ESRC Data Archive. An issue here is that while this was accepted and submitted on DVD, none of the material is available online from the Data Archive. Key issues affecting re-use also include the more complex requirements for obtaining informed consent from participants. A consent form developed with the ethics committee is used to request various permissions (e.g. to share with researchers, to show publicly) *after* recording. Storage is increasingly an issue especially as the current project uses high-definition video. With the availability of camcorders using disk-based storage rather than mini-DV tapes, the question of 'when to store what and where' is becoming more problematic and already means making ad-hoc use of external hard disks and unplanned-for demands on server storage. Web video sharing sites such as Vimeo offered some of the capabilities that were needed for sharing selected clips on a limited basis with colleagues and other researchers.

Robin Hill ~ Visual Cognition Group, Dept of Psychology ~ Eyetracking studies of gaze and perception of dynamic scenes.

Robin described experiments being conducted in the Visual Cognition Group using eye-tracking instruments, aiming to build corpora of videos and eyetracking data from people watching video clips. Surprisingly little previous work has been done, so an exploratory approach has been taken in a current Leverhume-funded project, using a wide range of video genres as cues. The sources being used for this include selected examples from the Augmented Multi-party Interaction (AMI) Project's corpus

(http://corpus.amiproject.org) of recorded meetings, instructional and demonstration videos, TV coverage of debates, and CCTV images. The corpora will feature differing aspects of attention to dynamic movements including faces during conversation, social gaze cues and gestures. Experiments mainly use compressed video clips as input, and as a result the video clips are overlayed with visualisations of participants' eye movements, developed in the research group to illustrate 'hot spots' being attended to. This data is derived from large quantities of eye position measurements taken at high frequency.

The main curation issues are, firstly; providing storage and access to large files (also bandwith issues here) initially for the project members and then also for colleagues across the research group. Secondly there is a need to provide wider access to interested academic parties, restricted according to the IPR limitations of the content. Thirdly a database structure and publicly accessible capabilities to browse the corpora are required. Fourthly, there are requirements to establish which video formats are appropriate for current needs and for posterity; and appropriate metadata standards and tools to help manage links between related files, and with tagging, coding and classification tasks.

Mark Hartswood ~ Social Informatics~ Video analysis of interaction as input to medical applications development

Mark discussed issues arising from using video across a number of projects in medical areas, usually to help evaluate prototypes of information systems. Video comprises a relatively small part of the research data for these evaluations, which normally employ a range of qualitative and quantitative sources. The social informatics team's work has increasingly focused on observational work where video has an important role.

Currently video is being used in a project to develop a training tool for radiologists learning how to screen mammograms. This builds on earlier work using video and still images, and demonstrating various ways that experienced radiographers physically work with radiographic film to decide the notable features of a mammogram. However field notes are more commonly used, especially where the setting is too sensitive for video recording. This has been the case for example in developing information systems for a psychiatric ward's handling of deliberate self-harm cases.

The medical application and sensitivity of the work presents the researchers with ethical issues of the intrusiveness of video recording, and practicalities of establishing informed consent where people (e.g. hospital staff) constantly move in and out of a setting. The high degree of oversight exercised through the NHS REC system has involved lengthy and protracted negotiations, in which researchers attempt to second guess the committee view and take a risk-averse position; e.g. data is typically destroyed at the end of projects.

The contextualisation needed to make video re-usable and portable is another issue. Although video seems to be in some respects a 'self-contained' source of data it never is completely; as can be seen in data sessions where the presenting researcher is often asked to clarify who or what is being shown. This becomes more problematic when sharing outside the research group, privacy allowing, as Mark illustrated with the still image of a mammogram annotated with (for example) several dates, the meaning of which had not been recorded.

Current developments to support Curation and Communication

Robin Rice ~ Data Library ~ 'Edinburgh DataShare from project to service'

Robin described recent work on the DataShare project. Initially funded as a two-year project to establish different exemplars of institutional data repository services across three institutions, this has received support to become a continuing facility for UoE researchers to submit and manage datasets they want to sustain[1]. DataShare envisaged the data repository roles in terms of several models; of a data sharing 'continuum', and of partnerships between researchers and curators throughout data and research lifecycles. In keeping with these the data repository will be supported through a range of web-pages and discipline-specific guidelines on data management, and training opportunities for researchers.

The requirements for these had also been identified through the JISC-funded project with DCC, the Data Audit Framework. Interviews with a broad range of departments had collected information about their data collections, and awareness, policies and practices in data curation and preservation, which highlighted the need for pragmatic assistance.

Nicola Hillhouse ~ Hillhouse Communications ~ Vidiowiki: Video as medium for interdisciplinary and public communication

Nicola gave an overview of Vidiowiki, a web-based platform for video presentations aimed at academic researchers. This provides researchers with capabilities to upload their short videos and presentations, and offers feedback on possible content relationships to other researchers' work. Currently in beta testing, the project's rationale comes partly from the 'compelling' nature of video and the enormous demand for it online. The core idea it aims to fulfil is the identification of 'cross overs' of concepts between research domains. The Vidiowiki system invites its users to create 'mind maps' of their work and draws connections using them. Features are being further developed based on feedback from the beta testing.

Angus Whyte ~ Digital Curation Centre ~ The curation lifecycle and video research data

Angus outlined how the SCARP case study on video data, had used the DCC Curation Lifecycle Model to organise interview questions to researchers on their research and curation practices. Considering the issues raised, it made more sense to think of several iterations of the curation lifecycle rather than attempt to fit a single one to the typical workflow in the projects involved. Three main cycles of curation were apparent; a '*planning and piloting'* phase that would aim to consider issues across the whole project, manage research material initially gathered and organised for individual analysis and some collaboration among the core research team. The main '*project curation*' phase would involve the bulk of data gathering, and some re-organisation of data as the nature and extent of it, and constraints affecting it's use, became clearer and it was worked on with more involvement of colleagues and peers. Then the '*long term curation*' phase began with moves to publish from the data and take more concrete steps towards archiving it.

The main issues identified included the uncertainties inherent in exploratory research approaches and exacerbated through technology changes and complexity. This lent

[1] The Edinburgh DataShare repository is available at: http://datashare.edina.ac.uk/dspace/

itself to a risk management approach to tackling curation risks. Working with video amplified some of the risks in areas of ethics, rights management and costs. It was difficult to identify the 'downstream' costs of video capture technology, and while the costs of video storage and capture were falling, costs of managing it were not, and it was getting more storage intensive. Storage decisions were a trade-off between space available, the size of the files, and sensitivity of their content- and therefore trust in the storage owner to observe the terms of consent. Dealing with analytic notes, content tags or descriptive information and technical metadata could be problematic if these were not exportable from tools used to share and access the material internally to those used for wider sharing with the project group and eventually online. Similarly with format choices, these were costly to change as transcoding video from the master or source to another format could take as long if not longer than to encode in the first place. Tools for online video asset management were specialised and changing frequently, and few met the likely need for browsable corpora that could continue to be annotated by the research community.

To address these issues there seemed to be roles for support services at School or Department and institutional or wider levels, beginning with advice in drawing up data plans and continuing at each phase with progressively more involvement of data repository and support services in the action taken. Managed storage, format migration/transcoding, access management, media streaming and repository management seemed likely candidates for support.

Discussion

Angus handed round copies of 'curation issues' that had been distributed before the workshop. The discussion referred to these indirectly and focused on the following-

Where is help most required with 'lifecycle management' issues?

Advice with data planning, formats, metadata and legal/ethical issues were seen as key needs. The 'data management plan' (e.g. ESRC) or 'technical annex' (AHRC) were necessary to get funding, and researchers would appreciate guidelines on what constitutes good practice, with examples of 'boiler plate' text that could be adapted. Help would be valued in keeping track of which formats are recommended, and which standards could be used to manage the relationships between file versions/ renditions.

Do 'lifecycle models' help plan a/v data management?

Comments were that they were generally helpful, but more difficult questions soon arose in applying them. Core questions were at what point video material gathered should be considered 'data', and what that term implied for curation. On the first point, some participants felt that recorded footage might be thought of as 'pre-data' until it was identified e.g. through capture and logging. This had important implications for the need for long-term curation, since this stage typically involved selecting around 10% of recorded footage for further work and while this selection might be revised during a project that would be unlikely afterwards.

A distinction was made between two senses of 'data' as the outcome of digitization and as evidence for research, and it was believed curation should take care of both senses. However it was felt that in some approaches, particularly survey-based methodologies, what constitutes 'data' is relatively well-accepted, while in others such as ethnographies this depends more on topic and perspective.

Software curation was also an issue. For example in a project assessing the benefits of a web-based tool providing streamed video and online annotation, the long-term accessibility and potential reuse of the annotations would depend entirely on being

able to run this tool together with the browser software originally used to make and view them.

What services will fit needs and how might they work between research groups, school-level and central services?

Advice was needed as already mentioned, and in the form of direct advice e.g. with budgeting, as well as more generic guidelines.

The collaborative aspects of video research were felt to be the most difficult in curation terms, to support intensive logging and tagging. Traditionally this has been the work of an individual, partly as in a lot of social science the 'lone researcher' has been the norm. Using web-based capabilities to work more collaboratively means that the volume of material that can practically be curated is self-limited to what a small group can work with, and readily understand the context of. Curation support at this stage needs a 'productive repository', so a corpus can be browsed and annotated, and those annotations become part of the curated object.

For long-term curation there was a need (beyond video data) for an institutional model; maybe reinventing the 'client-server' model to apply to data repositories. Something like a 'research executor' role was needed for collections of data whose owners had departed the institution, but the issue of context made this difficult.

What do researchers and service providers expect each other to do to preserve & curate video-based research?

The general view was the conversation needed to continue before wants and needs could be defined further. One suggestion was a website covering methodological as well as technical aspects of managing video data. ECDF or DCC might take up from where the workshop left off.

Who would researchers look to for help?

The Multimedia Team in Information Services could provide advice on formats, and were about to launch a podcasting service. Also the newly re-launched JISC Digital Media provided advice through their website and help desk. The forthcoming DCC case study for the SCARP project would also pull together sources of advice.