Archiving multi-media and web-based data: representation and sustainability

Louise Corti
UK Data Archive

IASSIST 2007
Multi-modal research

Data collections
- Text, numeric, audio-visual etc.

Products based on data:
- Research outputs
  - Publications, reports and resources
  - Substantive, methodological and technical publications
  - Educational resources
  - Tools and software
New forms of publishing

- Move from flat linear text-oriented outputs
- To web presentation of research results
  - guided tours through data – view audio-visual data
  - exemplars of data on web (as static web pages)
  - technical demonstrators to present data
  - front ends to explore and analyse data
  - Based on home-grown’ repository systems e.g. FEDORA
Web archiving

- Wayback machine enables static capture of web pages – 10 billion webpages from 1996
- Joint Web Archiving consortium in UK still are not dealing with interactive resources (eg back end databases)
- Web archiving is becoming an absolute necessity for data archivists
- UK Data Archive beginning to explore with some test cases
Representation of qualitative data

Four examples of innovation in presentation of research in the qualitative world:

- Access to browsing high profile qualitative data and research online
- A multi-modal ethnographic dataset
- A multi-modal longitudinal study of young people
- A digitised paper-based collection of conferences and group debates
Qualidata Online

- The interactive face of ESDS Qualidata
- moves beyond catalogue searching and data download to allow web-based free-text and filtered searching, browsing and retrieval of research data in real time.
- increasingly, data in the system includes not only traditional interview transcripts, but also audio and image files
- Based on XML marked-up data
ESDS Qualidata

Search interview transcripts across multiple data collections

This page enables you to search the available transcripts across multiple data collections (datasets). In addition to the keyword search, you may also apply any of the available filters (sex, year of birth, etc.) below. However, the site does not currently support using the filter only. That is, at least one keyword is required for any search.

Search for: [Input field]
Collections: [Dropdown: All]

Sex: [Dropdown]
Year of Birth: [Dropdown]
Geographical Region: [Input field]
Occupational Class: [Input field]

Do you want to search for any (OR) or all (AND) of your terms?

Default is "OR"
XML model transcript

- Representation based on style sheets with XML marked-up data using TEI
- Will utilise named entity mark-up to enable much more powerful searching
- New grant to develop a data exchange standard
  - Qual and quant (with help of ODaF)
- Using METS to describe whole collection of data and linkage between all possible objects and parts of objects
MIQDAS project

- MIQDAS has been exploring the methodological implications of archiving, distributing and sharing multimedia qualitative data sets.

- The project has been working with an existing multimedia data set in order to consider the interface between data sharing, research design and methodological decision making.

- Data generated and stored in a variety of media present particular challenges for effective, integrated and ethical archiving and potential reuse.
Ethnographic Hypermedia

- previous project created an “Ethnographic Hypermedia Environment” (EHE) based on the ethnographic study of a science centre

- EHE - a multimedia, hyperlinked presentation created on DVD using XHTML. Includes both:
  - representation – the ethnography
  - Multimedia data set – cleaned and hyperlinked
The Ethnography

This leads to the hypermedia dataset. It can be navigated freely. All the data referenced by the trails is contained here.

Resources

This includes links to related sites and a wide range of tools for interacting with the EHE as well as much of the collected data in non-hypertext form.

Reception

This is the main portal for the Ethnographic Hypermedia Environment (EHE)

The ethnography that is presented here is split into two "trails:"

- Physics and Physicality
- The Wow Factor

Physics and Physicality explores the social function of physical interaction at the research site. The Wow Factor is a sub-trail that expands on the use of spectacle in the performance of science communication.

These trails form the main body of the substantive work. To return to the main ethnography from any point in the EHE use the "Trails" link at the top of the screen.

If you are ready to interact with the EHE, then start here. Otherwise, choose "Getting Started."

Information

You are in the "reception" to the Ethnographic Hypermedia Environment (EHE). It serves as the 'home page' or 'portal' for this site.

The "Home" link in the grey bar at the top will return you to this page. You can also bookmark pages and zones of interest using your web browser as per normal.

Getting Started

Choose the section "Getting Started" for an introduction to the EHE. Once you are ready, the Trails zone provides the main substantive element. You can navigate directly to any of the zones by clicking the appropriate picture.

Help

The help pages provide information about how to best use the Ethnographic Hypermedia Environment.

Getting Started

For a set of guides about how to interact with the EHE, choose this link.

Brief Orientation

The EHE is split into 4 main zones: the ethnography (trails); the hypermedia dataset; the resources and the help files. Each zone is colour-coded to aid in navigation. You will also notice that each zone has a logo in the top right corner. In addition, when you put your mouse over a link the background colour will change to indicate what zone it links to.
Online Guide

- uses exemplars to illustrate four major dilemmas that researchers preparing their data for re-use need to consider how to:
  - ensure multimedia data can be contextualised
  - both methodological and substantive, and how best to describe it in relation to data generation issues
  - organise the data-set in a way that allows re-users to identify data records, understand the evidence-base and navigate through the data-set
  - tackle the ethical dimensions of deposition through considering the legal and regulatory frameworks as well as the ethical dilemmas of re-use
Making the Long View Project

- INVENTING ADULTHOODS data a unique QL study combining three ESRC-funded studies from 1996-2006

- Rich biographical accounts of all aspects of life of a diverse group of 100 young people (aged 11-17 in 1996) growing up at the turn of the 21st century

- Case Data:
  - Individual interview audio cassette, MP3 file and transcript
  - Questionnaire
  - Focus group audio cassette and transcript
  - Lifeline (& follow-up)
  - Memorybook
  - Case profile (first level analysis post-interview: processual features, substantive content, and researcher’s reflections)
Representing the data

- How to store qualitative data: by case, cross-sectionally or both

- Take a pragmatic approach to making technological decisions (e.g. digital audio file type), balancing quality and future-proofing with budget and storage space

- Clean and negotiate archiving consent as project goes but, in the QL context, leave anonymisation as late as possible

- Establish a standard page format and keep track of all anonymisation changes made in a standard format

- Save and back-up both cleaned and cleaned and anonymised copies of transcripts
Educational Evaluation project

- An electronic archive of significant educational evaluations representing a range of approaches and educational, political and methodological contexts

- A metadata vocabulary to describe the range of data, research instruments and techniques, interpretations, reports, documentation

- A range of technical tools designed to:
  - assist in data collection and description
  - ingress data and metadata into archives
  - disseminate contents
  - embed the archive in a Virtual Collaboration Environment - 'proxy' tool allows the archive to be used alongside a set of collaboration tools: chat, email archive, discussion, wiki, short-term file storage
Uses FEDORA

- Fedora Digital Repository Archive of educational research digital objects
- Open source platform enables a flexible and extensible collection structure, and exposure of records via the Open Archive Initiative (OAI)
- Can store a wide range of data types (audio, video, images, texts etc) while also separating data, metadata and any 'semantic' layers, and thereby enhancing exploratory and presentational power of the archive
- Allows not only resource description and discovery, but provides a basis for resources to be presented to the widest possible range of users through a range of web interfaces
FEDORA repository
Project Resource Browser

An Electronic Archive of Educational Evaluations

Repository

Home > Buddy > Explore the Archive

Buddy Evaluation: The Impact of Educational Technologies in the USA

Search Criteria

Search by: -- Browse ALL items --

Exclude items that only contain Metadata and no other Downloads

Search for Items

Search Results

Found 13 item(s) for the search criteria specified above.

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Item Category</th>
<th>Abstract</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buddy Contextual</td>
<td>Lane, Kathleen</td>
<td>Analysis, Interpretation and Judgement: reflexive account</td>
<td>In this excerpt from the interview, Cynthia Cole talks about the results of asking the members of the Waxman family to produce drawings of the family home into which a networked computer had been introduced. She describes some of the features of the children's drawings and contrasts their richness with the results of more formal interviews.</td>
<td>2006-04-19</td>
</tr>
<tr>
<td>Interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buddy Project Website</td>
<td>Cole, Cynthia</td>
<td>Impact and Engagement: website</td>
<td>The current website of the Buddy Project, with some information about the history and evolution of the project</td>
<td>1993-01 to 2006-01</td>
</tr>
</tbody>
</table>
Contextual interview

Bread and Dreams, a Case Study of Bilingual Schooling in the USA

Contextual Interview - 2006-02

Location of Interview: University of East Anglia, Norwich, UK
Interviewer: Elliott, John
Participant: MacDonald, Barry

Barry MacDonald discusses the case study of Bilingual Education in Boston, USA, funded by the Ford Foundation and published as 'Bread and Dreams'

Watch a clip from the interview

Download Full Transcript of Interview (RTF version)
Download Full Transcript of Interview (Text version)
Linking data and research outputs

Source to Output Repositories (StoRE Project)

- JISC funded under the Digital Repositories Programme
- addressing the area of interactions between output repositories of research publications and source repositories of primary research data.
- conducting user surveys to determine required functionality in repositories
  - for using primary data in source repositories
  - at the point of submitting to or downloading papers from output repositories
- general principles for middleware development to link source and output repositories together researched
- pilot demonstrator being developed using FEDORA (see Ken)
General move towards creating on-line resources to support research, learning and teaching:

- Economic and Social Data Service
- The Question Bank
- Research Methods Programme
- Researcher Development Initiative
- National Centre for Research Methods

- Digitisation projects
- Web-sites associated with ESRC funded research grants and programmes, as well as searchable on-line databases
- Learning resources based on data
On-line data related resources

- Development of an on-line resource is time-consuming and expensive and the full value of the resource only comes into play close to the point at which funding ends.

- Project, ‘Practical Exemplars on the Analysis of Surveys’ (PEAS, http://www.napier.ac.uk/depts/fhls/peas/) developed on-line resources to help survey analysts use weights and design factors.

- PEAS resources heavily used by researchers working with government surveys and similar datasets.

- Exemplars use several statistical packages (e.g., SPSS, STATS) and datasets drawn from major surveys.

- New versions of SPSS and STATA require updates to support materials and stability and sustainability are being considered.
Static or Ongoing?

One off-static resources:

- Archiving, such as via a data archive or learning repository, can provide a suitable solution for some resources.

- Ensures that deposited materials are appropriately documented, enabling future potential users to search for relevant content.

Many resources where static is not adequate:

- Solution may be to transfer materials requiring updating to a supported web architecture for ongoing support.

- Updating either by:
  - original team or commission externally-authored
  - or central technical support could be provided to maintain resources in the face of lack of end of project funding and changing web standards.

- Institutional repositories not yet set up to curate, support or maintain specialist resources.
Key considerations

- academic and technical review of potential resources for maintenance to ensure high quality and technical robustness – at time of awarding AND post project

- copyright and intellectual property rights, ethics and confidentiality of the resources

- citation...still no good system in place!

- work needs to be done on identifying underlying principles and examples of good practice
Options for research methods

Web service under proposal by NCRM:

- sustain resources with support from curating centre with a priority rating

- transfer into an archive such as Jorum, the Data Archive or open-source software repositories, such as OMII-UK (http://www.omii.ac.uk/)

- cease support by e.g. ESRC (possibility of self archiving)
Recognition that there are significant challenges and costs built in for long-term sustainability of rich and complex web-resources.

Design in representation also needs to be considered..... as does IPR and citation.

Need to consider longer-term maintenance of holistic resources.

Support linking data to outputs and back.

New technical solutions are moving ahead...fast.

Work done on non-proprietary data and file exchange formats will enable easier and cheaper portability and longer-term curation possibilities (less technical obsolescence...).

Web 2 might offer solutions – archives plus communications tools???