DP-DUG-01

AHRC – BBC Connected Histories Project

Minutes of the 1st meeting of Digital User Group held on 19 June 2017 at 11:00 in Room 108, Arts A, University of Sussex, Falmer.

1. Present

Tim Hitchcock (Chair) ; Adam Harwood; Alban Webb; Alex Butterworth; Bill Thompson; David Hendy; John Escolme; John Stack; Mahendra Mahey; Mike Dick; Peter Collier; Rob Cooper; Sharon Webb. Minutes: Denice Penrose

2. Introduction

The Committee Received the BBC Connected Histories Outline presented by Tim Hitchcock.

(a) This is a 5 year project with multiple partners
(b) The project will work with open linked data collections
(c) The project is underpinned by the BBC oral histories
(d) We intend to provide a popular web resource
(e) The project aims to establish a series of interventions for oral histories.
(f) This is the start of a long journey.

3) The Raw materials

Alban Webb provided a sample transcript from an interview of Wyndham Goldie.

(a) Project materials: The materials consist of 632 interviews with former members of BBC staff, and those external to the BBC associated with broadcasting in a range of formats. (See table below) The collection includes exit interviews with BBC staff, and commentary on national moments, such as the coronation. This raises the potential for a thematic approach to the materials, focussing on the role of broadcast media in the life of a nation.

Digibeta	121	CD	8
Beta SP	203	DAT	8
DVCam	104	Mag Tape	1206
DVD	18	VHS	284
MiniDV	8	35mm	over 500 reels

(b) Digitisation is currently underway on a tranche by tranche pattern, following annual oral history releases, with the current theme of Radio Re-invented. The accompanying materials have also been digitized.
    i) 2015: Election Broadcasting
    ii) 2016: The Birth of Television
    iii) 2017: Radio Reinvented
    iv) 2018: Britishness at Home and Abroad
    v) 2019: BBC and War: Hot and Cold
    vi) 2020: Entertaining the Nation
    vii) 2021: Inventing the Future: BBC & Technology

(c) The BBC post interview process involved the creation of a transcript, which was sent to the interviewees for approval. Embargoes were applied to sensitive material. The archive comprises of the correspondence related to the interviews, along with the transcripts and interviews.
(d) There are two main types of documentary evidence: the transcripts and the correspondence relating to the interviews (letters; payments made; permissions for use of interview etc.). This documentation is currently being digitized.
(e) A naming convention has been established for the material as follows:
    i) <File Reference>
    ii) <Interviewee Name>
    iii) <Interview Date>
    iv) <Transcript or Correspondence>
    v) <Page Number>
(f) The project team is currently investigating issues around data storage and the handling of project data.
(g) The materials consist not only of the BBC oral history collection, but also materials from the archives of Mass Observation and the British Entertainment History Project. The project will aim to create open linked data and standard navigation to compliment the collection.

The following points were among those made in discussion:

- There was discussion around the analysis and formatting of the transcripts. Some transcripts differ substantially from the interviews. It was suggested that an evaluation of these discrepancies could inform decisions around a strategy for analysis.
- The transcripts have been scanned as OCR, but include hand written notations, which cannot be easily digitized. The discussion considered whether it would be quicker to redo the transcripts by hand, or to use software. Transcriptorium (http://transcriptorium.eu/) is a program that could potentially be used to digitize hand writing, but would require 50 pages written in the same handwriting.
- There was discussion around open linked data based on the typed audio transcripts without the handwritten notes. The OCR extraction could be used for digital record entity extraction. Entity analysis can be done by running queries through DBpedia and then used to identify additional complementary resources. Genome can be also be used to cross check data http://genome.ch.bbc.co.uk/
- IT infrastructure for project was discussed. The project is part of Sussex Humanities Lab, a digital lab resource that can be used to some extent. As this is a partnership project we will be considering where amongst the partners we can identify best practice to emulate. The Sussex Humanities Lab is advertising for full time web develop. There is funding for a post to develop cataloguing / research fellow from April 2018. The project has access to university informatics & text analysis, and

there is funding to buy out time of professors. Digitization is being done by partner institutions.

.

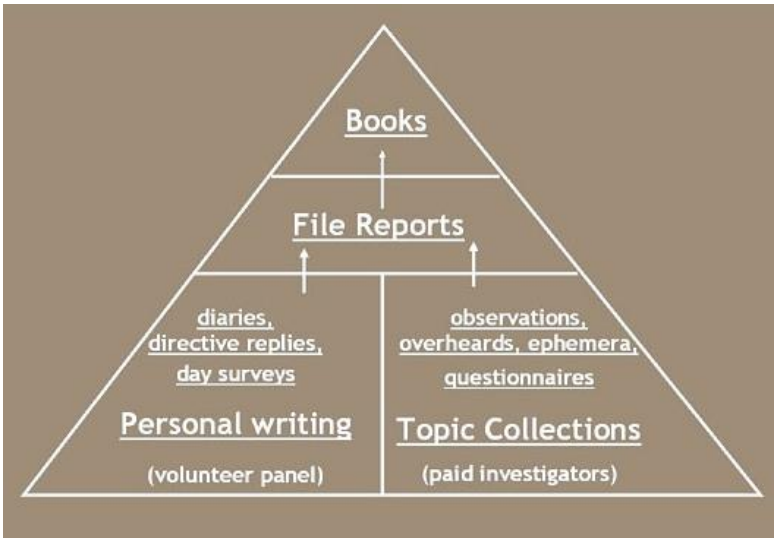
4) Technical plan

The committee received a copy of the detailed Technical Plan for the project, which was then discussed.

a) The development of an ontology was discussed, and the committee were invited to give input into the process, and to recommend any potential ontologies which could be utilised in the project. Recommendations were made to use a domain expert to read through the data and discuss the proposed data structures, and to make one individual responsible for the final ontology, with input given from others while building the ontology.
b) There was discussion around the use of a controlled vocabulary. The BBC has a core concept ontology list (http://www.bbc.co.uk/ontologies) and has a range of ontologies. The ontology used for the project will need to reflect academic scholarship as well as those used by partners such as Mass Observation.
c) A recommendation was made to identify potential users, and engage them in a process of identifying and tagging structures and formats, to ensure the data is user friendly. The data model will define the parameters of what is and what isn't coded. Establishing potential uses of the data will provide input into the data models which are needed. A range of workshops were suggested in order to facilitate this discussion.
d) When the new Research Fellow is recruited, it was recommended that the role of reading, listening and reviewing materials needs to be included in job description.

5) Digitalisation, and the current status of materials.

a) John Escolme outlined the BBC's current approach to the digitization of materials, using a tranche by tranche approach to make the process more manageable. 40-70 transfers are being done in each tranche, themed as per the list in 3b of these minutes. All of the transcripts for the oral history interviews have now been scanned, aside from those from the oral History of the North. These have a range of accompanying notes, and offer scope for a different approach to the interviews.
b) There was discussion around the standards used for recording and digitization, and identifying the appropriate format for preservation of audio and video files. The current industry standard is MXF. The existing digital formats meet the minimum criteria. It was recommended that the new interviews be recorded to the highest possible level to accommodate future developments, but it was agreed that .mov and .wav should be adequate for the current project
c) Mike Dick from the British Entertainment History Project gave a brief overview of the history of the project. The project was founded in 1986, and is run by volunteers. The archives include 700 interviews; 540 audio recordings, 160 video clips; 1200 audio cassettes and approximately 150 transcripts. An audit of the collection is needed to identify potential materials to include in the project.
d) Alban Webb gave an overview of the history of the Mass Observation Project and Archives. These provide information on audience reactions to broadcasts, and provide a rich source of data. An overview of the types of records included in the archive can be found on the next page.



The materials are predominantly hand written, which presents challenges for digitization. Adam Matthews have digitized some areas of the collection, and will allow up to 10% of the collection to be made openly available. More recent areas of the archive have not been digitized, which presents a challenge. Project funding includes buyout of Fiona Courage's time to participate in the project.

There was discussion around engaging with the data, and strategies for converting it into digital data. The British Library may have tools which could assist with this, but will need a sample selection to determine the viability of their tools for this project.

6) Data management

a) Adam Harwood outlined Sussex's storage structures for research data. It is anticipated that Figshare will be available at the end of October. This is a proprietary cloud based service to allow Sussex researchers to make their research public, and will be available for researchers or projects. Data on Figshare will automatically be transferred to Archivum, which keeps 3 copies of research data for up to 25 years
b) There was discussion around the immediate need to identify a storage space for materials, which will ensure that they are held with sufficient security to meet the confidentiality requirements, won't impinge any rights, and still allow access for the project researchers, or limited access for members of the Project Advisory Board.
c) In the longer term, decisions need to be made about the final storage of the data, including the new interviews, and how that is transferred to the library.
d) Sussex has existing guidelines for data storage, which need to be evaluated against the BBC's agreement for data management, in order to determine the most appropriate storage facilities.
e) A data management flow will need to be developed for the various aspects of the data.

7) Tagging and Encoding strategies

A brief discussion was held around Tagging and encoding strategies, with the recommendation that workshops be set up as soon as possible to begin working on strategies.

DP
27.06.17