
Proceedings Paper:
Hanchard, M. orcid.org/0000-0003-2460-8638 and Merrington, P. (2018) Developing a computational ontology from mixed-methods research: a workflow and Its challenges. In: Pitcher, L. and Pidd, M., (eds.) Proceedings of the Digital Humanities Congress 2018. Digital Humanities Congress 2018, 06-08 Sep 2018, Sheffield, UK. University of Sheffield , Sheffield

---

Proceedings of the Digital Humanities Congress 2018

Edited by Lana Pitcher and Michael Pidd

Developing a Computational Ontology from Mixed-Methods Research: A Workflow and Its Challenges

by Matthew Hanchard and Peter Merrington

1. Introduction

In this paper, we discuss four challenges we faced in developing a computational ontology (ontology hereon) from a mixed-methods social research project that examines the formation of film audiences. This includes a discussion of 'dual coding' in our qualitative data analysis (QDA), where we coded towards the development of a database and ontology at the same time as coding to develop theoretical concepts. We explain that the process of dual coding requires a rigorous coding strategy, which can be time-consuming and by extension generates a large volume of codes. In turn, the latter leads us to discuss data management as the second challenge that we encountered. As we explain in this paper, the large volume of codes generated through our QDA required us to be adaptable when working within the constraints of limited software availability. This need for adaptability was further extended into a third challenge that we encountered in working across different disciplines. Our QDA follows a qualitative social science tradition of

building theory inductively (ground-up) from the data. This conflicts with established practices in software development (e.g. building a database or ontology), which tend to build a structure, i.e. database tables, and then modify data to conform to that structure. In this paper we explain how we balanced the two opposing modes of working. As a final challenge, we note that our approach identified both 'overt' and 'latent' relationships between data. While overt relationships were specified during coding, the 'latent' relationships were only made visible through the ontology and provided a relational understanding in a way that traditional mixed-methods research could not. However, as we note below, making both types of relationship accessible in our research outputs raises several further challenges.

This paper is based on the early stages of a three-year Arts and Humanities Research Council (AHRC) funded project titled 'Beyond the Multiplex: Audiences for specialised film in English regions' (BtM hereon), grant reference: AH/P005780/1. BtM is a collaboration between the Universities of Glasgow, Sheffield, liverpool, and York (UKRI, 2017).[1] The project also works closely with the British Film Institute's (BFI) Film Audience Network (FAN). The project aims to develop a better understanding of how to enable a wider range of audiences to participate in a more diverse film culture. To do so, the project compares four English regions (North East, North West, South West, and Yorkshire and Humber) to generate insights into how film is consumed and understood regionally.

The research is organised into eight work packages (WPs):

- WP 1 – Development of a computational ontology as a tool to aid analysis.
- WP 2 – Analysis of 115 policy documents taken from key industry bodies to identify trends in film production and distribution.
- WP 3 – Development of a socio-cultural index to identify the factors that influence how and when specialised film audiences form, to compare their configurations within wider patterns of cultural consumption.
- WP 4 – Thematic analysis of 200 semi-structured interviews (50 per region) with audience members to explore how people engage with film (both mainstream and specialised), and to explore how such engagement features within wider patterns of cultural consumption.
- WP 5 – A survey with three waves (N=5000, n=500, and n=250 using a sample of 'within same-set' respondents) to capture details of film consumption over time.
- WP 6 – Thematic analysis of semi-structured 'expert interviews' with film industry policy makers and key stakeholders to explore the processes of production, distribution, and exhibition of specialised film.
- WP7 – Textual analysis of 16 film-elicitation groups (4 x per region) focusing on the interpretive labour undertaken (both by individuals and audiences) to

create meaning, and the cultural and interpretive resources that people draw on to do so.

- WP 8 – Development of data visualisation tools to make the ontology accessible and useable, both for its key stakeholders, academics interested in film consumption and audiences, and those working in the film exhibition sector.

Once data from all work packages is collated and coded, it will be parsed together and ingested into a graph database. At the moment, that will be MySQL-based, but we are considering other options such as SPARQL. The BtM project's planned outputs include a well-documented ontology, a socio-cultural index, a PHP-based public-facing website, and various publications, talks, and workshops. While the project explores regional patterns of film consumption (what people watch, when, where, and under what conditions), it also explores audience members' experiences and their meaning-making through film. In technical terms, this involves the development of an ontology 'ground-up' from unstructured qualitative data (interview transcripts) to formally describe film audiences and specialised film consumption, along with the formal elements and attributes of film. We started by developing an initial draft of an ontology, and then used that draft version as a baseline to inform a set of 'initial themes' that we later drew on within our QDA. In return, as we coded interview transcripts, our QDA informed an ongoing iterative revision and modification of the ontology. This two-way informance (between QDA and ontology development) presents a challenge in balancing the structure required to develop both a database and an ontology with the inductive process of coding unstructured data in QDA. In the next section, this paper

provides a contextual background to the research. It then moves onto a discussion of our workflow, our development of the ontology, and our nuanced use of QSR NVivo for the QDA.

2. Background Context: A New Approach to Exploring Specialised Film

While access to mainstream film is generally good across England, access to specialised film is low in regions outside London (Dickinson and Harvey, 2005) which limits opportunities for people to experience a diverse film culture outside of the capital. Following the BFI's definition, 'specialised film' encompasses films that sit outside any mainstream or highly commercial genre including small-scale British films, foreign language films, feature-length documentaries, artists' films, archival films, classic cinema, and films with unconventional narratives, themes or cinematic techniques (BFI, 2017).

In 2016, there were 526 specialised films released in the UK. This made up 64% of all film releases, but it only accounted for 3% of total box office receipts (BFI, 2017). At present, there is a need to know more about the specific audience(s) of specialised film in order to effectively develop future national policies, particularly from a regional perspective.

Exploring regional film provision demonstrates the variances in the way specialised film is accessible and consumed across England. In 2016 there were 4,150 cinema screens in the UK in 766 cinema venues – 104 more than in 2015 (BFI, 2017). North East England had the lowest provision, with 5.2 cinema screens per 100,000 people (BFI, 2017). By comparison, South West England and London had

6.0 and 6.7 cinema screens per 100,000 people respectively (BFI, 2017). Alongside provision of cinema screens, there is a regionalised disparity in cinema admissions too. For example, in 2016, cinemas in London saw an average of 3 admissions per person over the year (BFI, 2017). By contrast, in the North East the annual average for the same year was only 2.4 (BFI, 2017). There is a similar disparity in the diversity of screen content too; of the 7% of UK cinema venues that primarily focus on showing specialised film, a third are located in London while only 2% are in the North East (BFI, 2017). The BtM project aims to understand the implications of these disparities in more detail. Developing a computational ontology and pairing it with mixed-methods research allows us to work towards a robust analysis that integrates micro-sociological accounts of individual specialised film-related practices and experiences with larger-scale aggregated data.

Film audience research tends to draw on either audience surveys and box office data collection to describe audience figures, or small-scale qualitative studies focussed on audience experience at particular events or venues. However, neither of these approaches can fully address how audiences interact with, and relate to, various types of film, or how people draw on their social and cultural resources within the filmic practices and experiences. In addition, neither approach sufficiently addresses the industry and policy contexts in which such audience interactions take place. To address these shortfalls, the project combines both large-scale (through the survey) and small-scale (through interviews and film elicitation groups) approaches to data gathering and includes an analysis of the policy context. Rather than forcing a theoretical framework onto the various data as a means of bringing them together, we approach the multiple datasets through a digital humanities approach in

order to identify relationships between datasets, and between audiences and specialised film.

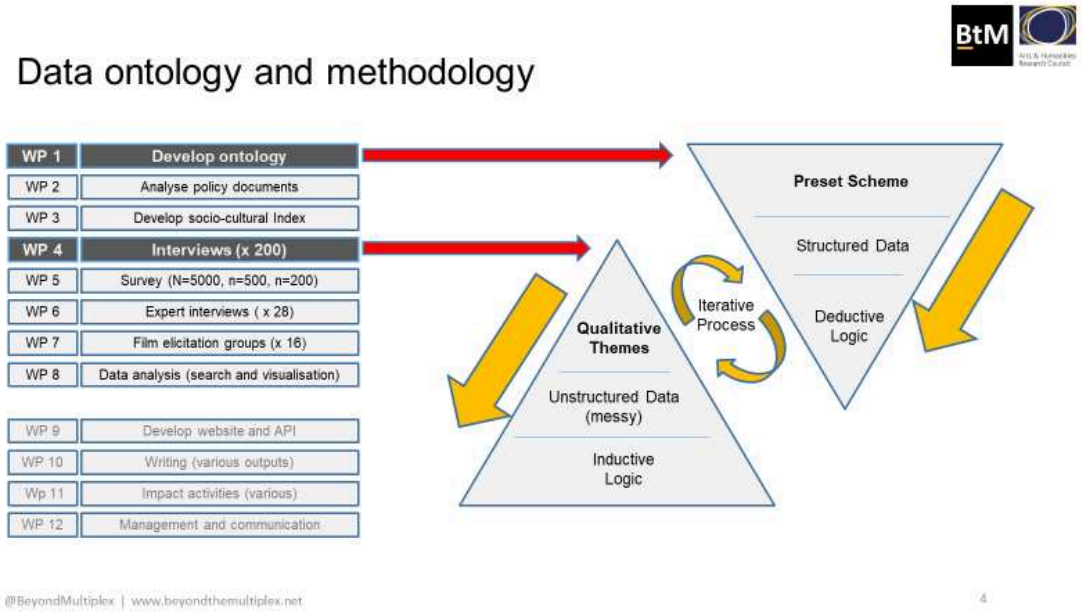
To frame our understanding, we follow Livingstone (1998), who suggests that audiences should be conceptualised as relational and interactive; an approach that acknowledges the diverse sets of relationships between people and media forms. For example, we do not treat cinema audiences abstractly – as a group of individuals that are geographically and temporally bounded within a specific place (the cinema space) at a specific time (the film showing). Instead, we consider their experiences of being brought together, their rationale for viewing a specific film at a specific time and place, their practices of selecting, watching, and sharing films, and their relationships to one another. Throughout our data gathering and analysis, we remain sensitive to an understanding that film audience members' practices and experiences are subject to constant change, for example through the introduction of platforms for online film consumption and distribution. Approaching film audiences as relational and interactive means asking how specialised films are located and understood as part of people's social and cultural practices, and not in isolation. In turn, this emphasises a need to focus on the modes of connection, relationship, and communication that make up the development of what Livingstone (1998) calls 'audiencehood'. Doing so helps us to identify the types of audiences that configure within specific contexts, and to explore how they do so. However, those types of insights are not readily accessible through either large-scale quantitative study (e.g. analyses of box office data) or small-scale qualitative studies of audience experiences (e.g. interviews with cinema goers). Therefore, the project employs a mixed-methods approach to draw together research at different scales for an

integrated analysis, using an ontology to understand the interrelations between processes of production, distribution, and consumption.

3. Workflow: Combining Mixed-Methods Research with a Computational Ontology

Typically, developing a data model, database, or computational ontology follows a well-planned and documented linear process. In short, developing an ontology allows researchers to map and classify the "*...components and characteristics of a particular knowledge domain...*" (Pidd and Rodgers, 2018). However, this process tends to be steeped in a deductive logic. That is, a developer builds a data model or set of tables, sets up queries and reports, and builds relationships between tables. This provides a structure which data can then be tidied, organised, and modified to fit. By contrast, qualitative data analysis (QDA) is messy, unstructured, asynchronous, and requires the overarching structure to emerge inductively through analysis rather than being forced onto data a priori. For QDA of interview transcripts especially, the natural flow of conversation rarely follows any neatly defined or taxonomically pre-set order (Edwards and Holland, 2013) – at least not when people talk about their past experiences of watching films. To that end, while the ontology and database both require data to be structured, for the QDA it was important, epistemically, to maintain the emerging structure of analysis. Thus, any claims we make about audience experiences must be based on participants' accounts and not on a reflection of how well we made their accounts fit a pre-existing scheme, itself laden with personal preconceptions and value-judgements. In this section, we discuss how we balanced those two opposing logics.

Figure 1: Balancing inductive and deductive logics through WP4 QDA and development of the ontology



At the start of our QDA, we drew on a pilot project (Corbett et al., 2014) to identify a set of themes that we felt were likely to emerge from the data. We organised those themes into a basic scheme composed of an initial set of entities with classifications below them. We used those initial entities and classifications as a baseline for developing the database and ontology, and as our initial themes for coding the WP4 interview transcripts. However, we remained open to an understanding that the QDA would evolve the structure of our entities and classifications.

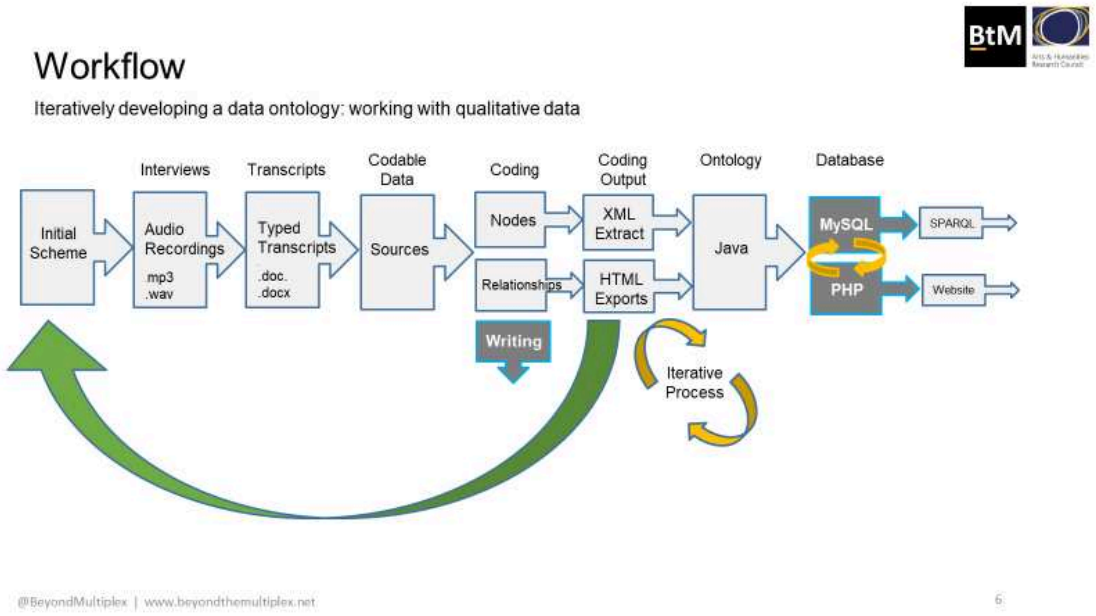
After drawing up an initial set of entities and classifications, we conducted the 200 semi-structured interviews, and input transcripts of interview recordings into QSR NVivo as text documents (as Sources in NVivo's terminology). As a market-leading software dedicated to mixed-methods and qualitative data

analysis, NVivo handles various file types, ranging from video, social media outputs, survey data, and text-based documents (NVivo, 2019). Each Project (NVivo's term for their .nvp file type) is a standalone relational database that allows users to code documents on-screen through well-designed graphical user interface (GUI) as a 'front-end'. The GUI enables researchers to code by assigning words, sentences, paragraphs, or any portion of a Source to one or more Nodes (NVivo's term for a text-based label or qualitative 'code'). Typically, in grounded theory or thematic analysis, researchers merge their Nodes, move them about, and sort them into a narrowed hierarchical scheme as the analysis goes on, leading them to identify specific concepts and categories that have emerged from the data (Charmaz, 2006, 2015). Alongside coding through the development of Nodes, NVivo also allows researchers to create Relationships (NVivo's term) between two Nodes, and to label or name that Relationship Type (another NVivo term). The latter is an aspect that is well-suited to the development of a graph database, e.g. with Nodes tabled separately as subjects and objects (or entities and entity characteristics), Relationship Types can be stored in a link table as predicates.

It is the Nodes and Relationships we developed in NVivo that have allowed us to generate theory about people's experiences of watching film and being part of an audience, and to describe their patterns of consumption. In this, the explanatory power of our QDA has been extended by combining the theory developed in WP4 with data from across the other work packages holistically. For example, the survey findings (primarily via cluster analysis and latent class analysis) have allowed us to further explore and classify the demographic makeup of audience types identified in the QDA.

As an ongoing process, we regularly extract all the Nodes in our Project (including their position within the coding scheme) as a single XML file. We also export all Relationships as HTML files. At midway through our analysis of the 200 interviews, we had over 3,300 Nodes and 622 Relationships. To work with an Extract and set of Exports of that size and scale, we parsed them together as text in XML, and then built the database in Java before storing it as MySQL. In this, Java offered no specific benefit to our workflow; it was selected simply because our developer was familiar with its syntax, and other languages may have been used instead. Later, we may move from MySQL to SPARQL to store the database. At the front-end, we plan to develop a PHP-based public-facing website, drawing on open-source visualisation tools (e.g. Leaflet) for user-friendly mapping and data visualisation.

Figure 2: Overview of our workflow for developing the ontology



In developing our workflow, we found that close collaboration between qualitative social researchers and digital humanities-focussed software developers was essential in balancing such an iterative approach. That is, the workflow required an approach that respected both the deductive logic required to develop a database and ontology (e.g. by making data fit a pre-existing structure) and the inductive logic required for developing theory from unstructured data without sacrificing participants' voices or forcing an a priori structure onto the data. To that end, our workflow was aligned to what we see as the core aim of digital humanities – to enable original sources to speak for themselves via digital means, without a researcher or developer imposing their personal values or biases onto the output data product.

4. Challenges: Issues Encountered and Ways of Working around Them

In following the workflow described above (Figure 2), we encountered several challenges; in particular, coding with a database and ontology in mind at the same time as coding to develop theoretical concepts – a process we call 'dual coding'. Generally, QDAs either start with no pre-existing concepts and build a coding scheme completely inductively to develop a theory that is grounded in the data, e.g. grounded theory (Charmaz, 2006), or they start with a few themes that are then interrogated and refined inductively through the analysis of data, e.g. thematic analysis (Silverman, 2010). While both lead researchers to make claims about the data, the latter draws on claims that are steeped and grounded within a close reading of source material.

As noted above, in our workflow we took a set of initial entities and classifications as our starting set of themes and developed a hierarchical coding scheme of Nodes and Relationships from them through a close reading of each Source. By contrast, tagging or marking-up documents for ingestion into a database (e.g. XML tagging) typically involves a systematic approach aligned to a pre-set scheme. We felt that doing so with qualitative data would lead to deductive verification of the pre-existing scheme, and thus reproduce it without generating any analytical insight. Balancing the two different approaches and coding the data towards both simultaneously was a key challenge. On the one hand, the coding scheme needed to be neatly hierarchical and consistent. For example, when participants discussed watching the feature film 'Return of the Jedi', we could code their discussion to the Node called 'Star Wars', sat hierarchically under another Node called 'Film series'. We could also code it to a specific Node called 'Return of the Jedi', sat underneath 'Star Wars'. In turn, this would enable the film to be recognised as both a specific standalone film, and as part of specific film series. However, on the other hand, we found that people rarely speak so taxonomically in the flow of natural conversation, e.g. in interviews (Edwards and Holland, 2013). Instead, when asked to discuss their favourite film, participants often used phrases like *"I love Star Wars..."* (Sarah). In this, it is not clear whether Sarah is referring to the 1977 original title (Star Wars Episode IV: A New Hope), the most recent release of the Star Wars series (Star Wars Episode VIII: The Last Jedi), or another film in the series. Likewise, the quote provides no way for us to know whether the participant is referring to a film or series specifically, or if they are using the term 'Star Wars' as shorthand marker for all space-based sci-fi operas. To counter such ambiguities, we created Nodes for each possibility. Following the example above, we created a Node

for '*Star Wars [Generic]*' to covers its wider use, alongside a Node for '*Star Wars [series]*' to cover specific instances where participants explicitly discussed the Star Wars film series itself. We also created Nodes for each film title in the series. Coding in this way enabled us to maintain consistency for the ontology, whilst providing enough flexibility for the development of concepts and theory. However, we found that it took a lot of time to 'dual code' in this way. It also generated a large volume of Nodes and Relationships.

In turn, the large volume of codes generated by the QDA raised challenges for data management. To clarify, one of the practical limitations we encountered throughout our research was software limitations. This stemmed from licensing constraints. While it is often omitted from methodological discussion, real-world research tends to be delimited by the software licences and licensing models available to an organisation or institution. As such, researchers often need to adapt and work around such practical constraints. In our case, the Universities that our project team members work at have all purchased enterprise (site-wide) licenses for QSR NVivo 11 Pro, rather than the Server version of NVivo 11 or NVivo 12 For Teams. While the latter two options allow multiple users to work simultaneously on the same Project, QSR NVivo 11 Pro requires that researchers work separately on separate standalone Projects and then merge their Projects together manually as .ldf files (Hanchard, 2019). With multiple researchers coding independently on a Project that boasts over 3,300 Nodes and 622 Relationships, regularly merging Projects allowed us to maintain a degree of consistency as a workaround, with the master version of all merged Projects discussed and amended collaboratively with each staged iteration. In turn, we found that coding in this way raised several issues: different researchers generated duplicate Nodes to mean different

things, but with the same name – an issue we rectified by renaming one of them; Nodes that represent the same concepts or category were generated in two or more places – an issue we rectified by merging the Nodes in one location; similarly, at times renamed Nodes and their original counterpart both existed within the Project – which we rectified by merging the old Node into the new one (to retain any new coding towards the old Node); at other times we found Relationships that referred to Nodes that had since changed e.g. that had been renamed or merged into other Nodes – an issue we rectified by manually tracing back the original Node and redefining the Relationship to account for such changes; we also found that some of the Nodes that we had generated intuitively were better placed as Relationships, leading us to modify them. For example, we changed a Node labelled '*Being part of an audience… depends on where you watch*' into a Relationship labelled '*Part of Audience (CHANGES WITH) Where watched*', to better suit the needs of the database and ontology.

Addressing these challenges led to several changes in the database and ontology structure. While ongoing iteration and revision of the coding scheme is a typical aspect of QDA (Bazeley and Jackson, 2013), changes to the coding scheme drawn on to develop a database or ontology means that the scheme upon which the database tables are built are subject to constant change; parts disappear, re-appear elsewhere, and move under or above other parts, and as such they provide little or no stability tor developers to work from. As a compromise, we agreed to work towards a hierarchy, where the highest two levels of Nodes identified part way through the QDA would be stabilised, while lower levels would be fleshed-out through further coding. When around a third of all interviews had been coded, we agreed to maintain the first two levels of the coding

scheme. However, we also agreed to remain open to the possibility that Nodes at those top-levels may need to change and noted that we would have full flexibility on how the QDA and Nodes beneath those two levels would be revised.

As a fourth challenge, following the workflow described above led us to identify both 'overt' and 'latent' Relationships between the Nodes in our data. To clarify, when we coded a portion of a transcript and assigned it to a named Relationship between two Nodes (and a Relationship Type), the Export of that Relationship could be easily converted into a subject, predicate, and object for use in our development of a graph database and in our development of an ontology. It is an 'overt' relationship because we named and assigned each element (Nodes, Relationship, and Relationship Type). However, by coding to Nodes as part of our QDA, we often coded the same part of a Source (interview transcript) to two or more Nodes. In the Star Wars example above, we may have coded a portion of text in an interview transcript to both a film title, a film series, and a specific sub-genre of films. In doing so, we generated a set of 'latent' relationships between Nodes – intersections of Source content coded to two or more Nodes. Latent relationships appear within the Extract and provide a ready means for us to explore how such intersecting Nodes relate to one another, but they do not provide any further detail on the Relationship itself (e.g. *how* the Nodes relate to one another). While we are working on a solution to identify latent relationships in our database, they will be accessible and easily visualised through the ontology.

5. Conclusion

This paper has described the design of our mixed-methods research, our development of an associated computational ontology, and the workflow followed in both. It has also discussed the challenges we encountered throughout our workflow, and the ways we worked around them. For example, we described the issues for data management that dual coding presents, and the need to work adaptively and collaboratively to work around them. We also described a strategy for working around a specific limitation in access to software (and licences). In addition, we highlighted how an interdisciplinary team from opposing theoretical paradigms can work collaboratively within a digital humanities context to develop computational tools (e.g. a database and a computational ontology) without sacrificing the epistemic rigour of qualitative social science and its inductive generation of theory. As a final challenge, we noted that the approach we have taken in this research serves to identify both overt and latent relationships between data, leading to a richer analysis than mixed-methods research could achieve alone. To that end, we hope that our workflow, the challenges described in this paper, and our resolutions of them, might be illustrative in helping future digital humanities scholars to continue enabling the original sources of data to speak for themselves via digital means, without imposing personal values or biases onto them in the data products and theories that they develop.

6. References

Bazeley, P. and Jackson, K. (2013) *Qualitative Data Analysis with NVivo*. London: Sage.

BFI (2017) *BFI Statistical Yearbook 2017: Film Forever.* Report. London: BFI Research & Statistics Unit.

BFI (2018) *BFI Statistical Yearbook 2018: Distribution and Exhibition*. Report, London: BFI Research & Statistics Unit.

Charmaz, K. (2006) *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. London: Sage Publications.

Charmaz, K. (2015) 'Teaching Theory Construction with Initial Grounded Theory Tools', *Qualitative Health Research*, 25(12), pp. 1610–1622. doi: 10.1177/1049732315613982.

Corbett, S., Wessels, B., Forrest, D., and Pidd, M. (2014) *How Audiences Form: Exploring Film Provision and Participation in the North of England*, Available at: https://www.showroomworkstation.org.uk/media/FilmHubNorth/How_Audiences_Form_Full_Report_UPDATED.pdf (Accessed: 26-Feb-2019).

Dickinson, M. and Harvey, S. (2005) 'Film Policy in the United Kingdom: New Labour at the Movies', *The Political Quarterly*, 76(3), pp. 420–429. doi: 10.1111/j.1467-923X.2005.00701.x

Edwards, R. and Holland, J. (2013) 'What forms can qualitative interviews take?', in Crow, G. (ed.) *What is Qualitative Interviewing?* London: Bloomsbury. pp. 29–42. doi: 10.5040/9781472545244.

Hanchard, M. (2019) *Using NVivo to structure a computational ontology*, Available at: https://www.qsrinternational.com/nvivo/nvivo-community/the-

nvivo-blog/using-nvivo-to-structure-a-computational-ontology (Accessed: 26-Feb-2019)

Livingstone, S. (1998) 'Audience research at the crossroads, the 'implied' audience in media theory', *European Journal of Cultural Studies*, 1(2), pp. 123–217. doi.org/10.1177/136754949800100203

NVivo (2019) *What is NVivo*, Available at: https://www.qsrinternational.com/nvivo/what-is-nvivo (Accessed: 26-Feb-2019)

Pidd, M. and Rodgers, K. (2018) *Why use an ontology? Mixed methods produce mixed data*, Available at: https://www.beyondthemultiplex.net/why-use-an-ontology-mixed-methods-produce-mixed-data/ (Accessed: 26-Feb-2019).

Silverman, D. (2010) *Doing qualitative research: a practical handbook*. London: Sage.

UKRI (2017) *Beyond the Multiplex: Audiences for Specialised Film in English Regions, UKRI gateway to publicly funded research and innovation*. Available at: https://gtr.ukri.org/projects?ref=AH%2FP005780%2F1 (Accessed: 26-Feb-2019).

1. BtM team members: Bridgette Wessels (PI), David Forrest, Andrew Higson, Michael Pidd, Simeon Yates, Matthew Hanchard, Peter Merrington, Katherine Rogers, Roderik Smits, and Nathan Townsend. ↵