Participations
Journal of Audience & Reception Studies

Volume 19, Issue 2
March 2023

Analysing film audiences: the merits of a data ontology

Michael Pidd
University of Sheffield, UK

Abstract

A data ontology (sometimes referred to as a *web ontology*, or a *computational ontology*) is a type of data model which is used to describe a knowledge domain. It enables digitised primary source content to be structured, stored and analysed in accordance with the characteristics of the domain, thereby overcoming the constraints of the primary source's medium and format. For example, in the project Beyond the Multiplex (2017–2021) a data ontology was used to model the domain of film audiences. It enabled the contents of 227 interviews, 16 film elicitation workshops, 114 industry and policy documents, and a survey of over 5,000 participants to be structured, stored and analysed holistically, irrespective of their different media and formats. However, most projects in the arts, humanities and social sciences that have used data ontologies have been influenced by science and engineering, where ontologies are formal categorisations of physical, measurable things. In contrast, the Beyond the Multiplex project developed a data ontology that models abstract concepts such as how people think, feel, experience and remember film, in addition to physical, measurable things such as film distribution and exhibition. This approach underpinned theoretically informed empirical research into how film audiences form as a social and cultural phenomenon.

In this paper we will explore the data ontology of film audiences that was developed by Beyond the Multiplex, and the effect that this model had on our interpretation of the project's primary sources. In particular, we will examine the ontology's merits as a tool for film audience research: did it deliver new insights?

Keywords: film, audiences, data, ontology

Introduction

In 2015 at a branch of *Starbucks* next to the University of Sheffield, a discussion about a project funding bid led to the idea that a thing called a *data ontology* could be used as a novel way of annotating and interpreting diverse types of primary source data. In the course of trying to persuade the different stakeholders who would be necessary to realise such a project, we evangelically proclaimed the value of this approach, but sometimes struggled to articulate what a data ontology actually *is*. The concept is novel in the arts, humanities and social sciences, and possibly completely unknown in parts of the creative industries. Eventually we simply reassured stakeholders that this was *a special type of database*, even though this description is wholly inaccurate.

The project was eventually funded by the Arts & Humanities Research Council (AHRC) in 2017, called Beyond the Multiplex: Audiences for Specialised Film in English Regions, and led by the University of Glasgow with the universities of Sheffield, York and Liverpool.[1] Four years later, in 2021, the project's principal digital output was published, a data platform that provides access to a large dataset of diverse primary sources, underpinned by the data ontology.[2] An online practical guide to using data ontologies in the arts, humanities and social sciences was subsequently made available (Pidd, 2021). Then in 2023 the project's research findings were published, *Film Audiences: Personal Journeys With Film* (Wessels et al., 2023).

With the completion of the project, this paper reflects on what a data ontology is, how it can be used, and what its value is for researchers in the arts, humanities and social sciences, by drawing on the project's experience of developing and using one. This is a frank account of why we used a data ontology, how we used it, and what the results were. Hopefully, the reader will be left at least with an understanding that an ontology is not a special type of database – it is much more special than that.

What is a data ontology?

In information science a *data ontology* is a formal description of a knowledge domain or subject area. It is sometimes referred to as a *web ontology*, *computational ontology*, or simply *ontology*. Whereas ontology as a broad philosophical discipline is the study of 'being' and of the 'existence of things', data ontologies categorise and model the qualities and relationships of things that exist. Importantly, a data ontology describes a knowledge domain that is represented and communicated using digital data. This data can be in the form of human discourse – formal and informal texts, including books, social media, and email, as well as video and audio. It can be data that has already been categorised or

[1] See https://gtr.ukri.org/projects?ref=AH%2FP005780%2F1 (2017–2018) and https://gtr.ukri.org/project/F73C3745-FF69-49B1-BD46-A0960082E90A (2018–2021).

[2] See https://www.beyondthemultiplex.org

labelled in some way, such as spreadsheets, databases, and surveys. Or it can be mechanical data, such as signals and programming logic. The Disease Ontology[3], for example, describes the domain of disease from a biomedical perspective so that all publications about disease are able to reference the same concepts, helping to ensure that scientific discourse about disease is unambiguous, and a single disease concept can be located across multiple publications within a repository.

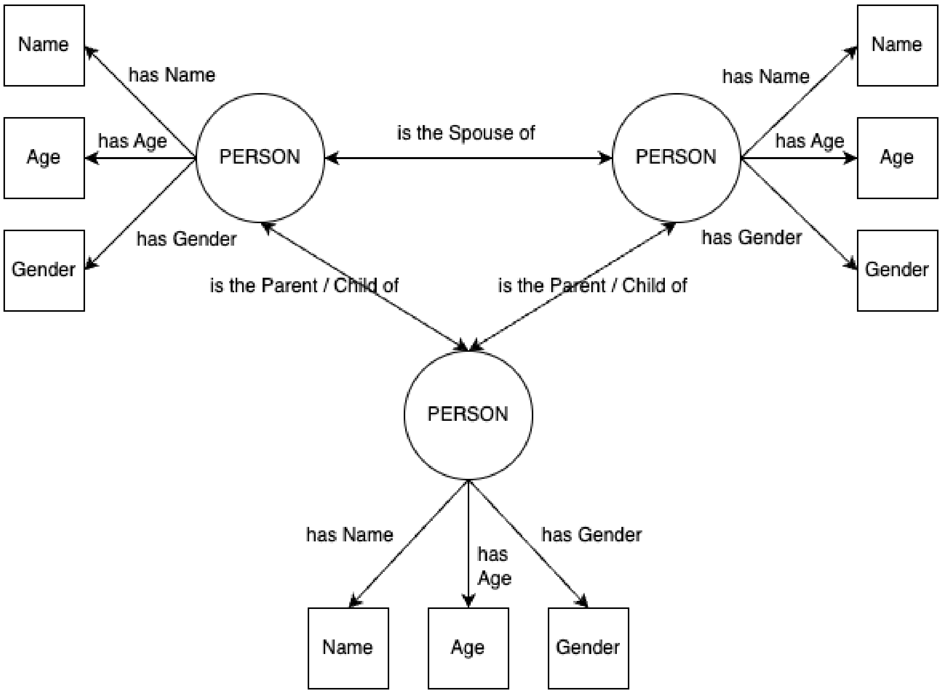
Data ontologies vary considerably in their scope and complexity, but most share the following characteristics: they have *concepts* (sometimes called classes, objects, or entities) which are things that exist in the knowledge domain; each concept has one or more *properties* (sometimes called attributes) which are the characteristics of the thing; concepts have *relationships* between one another to describe how things interact within the knowledge domain; and many ontologies have *rules* that describe how the things are permitted to exist and interact.

For example, a family tree would be a familiar ontology for describing kinship as a knowledge domain. There are people (concepts) with properties, such as name, gender, date of birth and date of death. The people have relationships with one another, and these relationships have properties, such as ‘father of’, ‘daughter of’ and ‘sibling to’. Each person can have multiple and different relationships, but there are rules that govern these. A person cannot be both the father of and the spouse of the same person. Likewise, a person cannot have a date of birth that occurs after their date of death. All of this is obvious to us, but these structures and the rules that govern them are important for modelling knowledge in a way that is interpretable for a computer, because a computer only sees data and not their real-world meanings and rules.

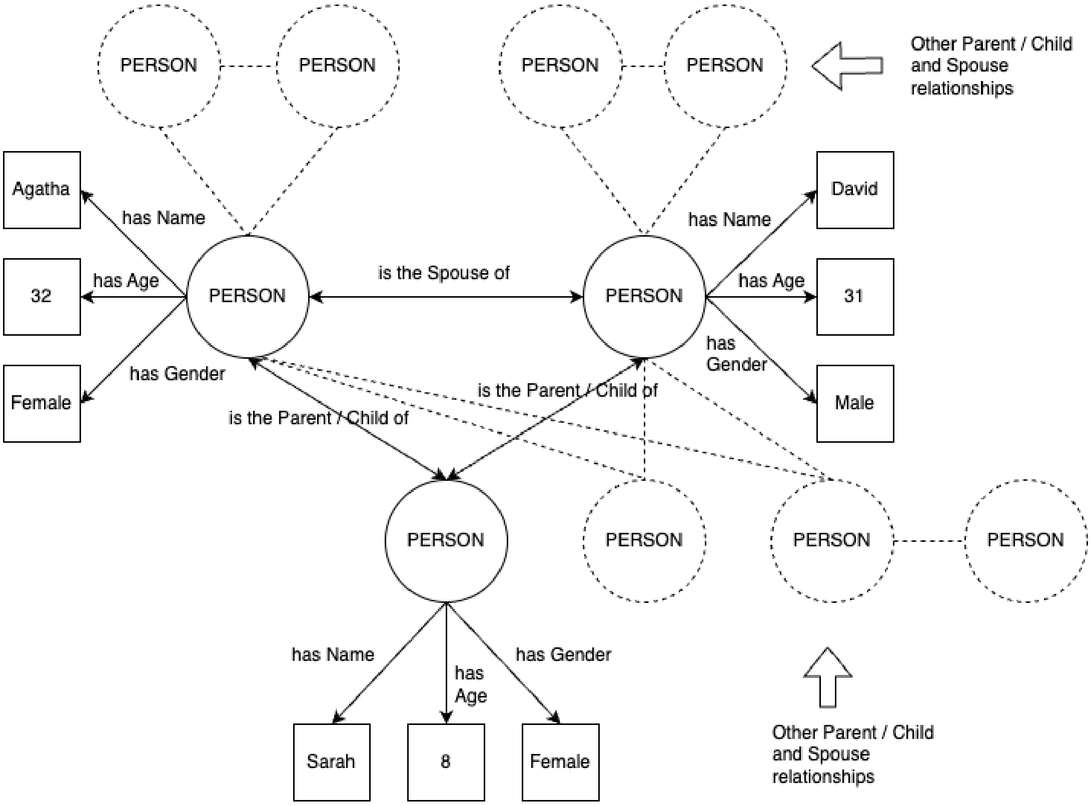
Another important characteristic of data ontologies is that they are not the data itself. An ontology is a *data model* in the sense that it is can be imposed onto existing data, as a metalayer, to standardise and describe concepts that are referred or inferred within it; or it can be used as a model when creating new data, controlling what the new data can be and how it must be structured.

[3] See https://disease-ontology.org

In order to illustrate this, here is our kinship ontology represented diagrammatically:



And here is what some data might look like if we structure it using our kinship ontology:



Clearly, we can use data ontologies to describe lots of knowledge domains that lend themselves to categorisation and systematisation, such as genealogy, biology, medicine, engineering, finance and linguistics. As such, most published ontologies address these types of domains, such as the Gene Ontology[4], the Financial Industry Business Ontology[5], the Geopolitical Ontology[6], and the Foundational Model of Anatomy.[7] Less common are data ontologies that deal with socio-cultural phenomena and ideas, perhaps because these knowledge domains are harder to model due to their variability and abstractness, or because the value of ontologies is less well understood within qualitative research practice.

Apart from linguistic databases that share some of the qualities of data ontologies, such as WordNet[8], ontologies in socio-cultural domains have tended to focus on physical, measurable things: people, places, documents, artworks, buildings, events, and time. CLARIAH's *Awesome Ontologies for Digital Humanities*[9], for example, includes general ontologies that describe physical, measurable things within the humanities broadly, as well as within specific knowledge domains such as history, musicology, language and literature. This focus on physical, measurable things within the humanities is perhaps because the majority of these data ontologies have tended to be created and used by information management professionals, concerned with describing archives of objects to improve resource discovery. The CIDOC Conceptual Reference Model (CRM for short)[10], for example, is intended to describe cultural heritage objects and collections for the benefit of galleries, libraries, archives and museums. Cultural heritage is extremely diverse, and this is reflected in the size and detail of CRM. CRM is a model of concepts and relationships so detailed and comprehensive that at first inspection it seems capable of describing virtually everything related to human culture, and yet it does not model ideas – which underpin all cultural heritage. Of course, there are lots of physical, measurable things in the human world, present and past. But ideas tend to be the driving force behind most of them. Only the Literary Theme Ontology[11], included in CLARIAH's directory and which is designed to describe themes in literary fiction, attempts to model aspects of human emotion and intellect rather than physical, measurable things.

A film audience ontology

The film audience ontology developed by Beyond the Multiplex has similar ambitions to the Literary Theme Ontology – it seeks to describe the emotions, ideas, contexts and experiences of viewing film, in addition to film's physical manifestations as works of art, people, places, distribution and exhibition. As such, it is very different to virtually all other published ontologies.

[4] See http://geneontology.org
[5] See https://spec.edmcouncil.org/fibo/ontology
[6] Developed by the United Nations' Food and Agriculture Organization, its status and existence is now unclear.
[7] See http://si.washington.edu/projects/fma
[8] See https://wordnet.princeton.edu
[9] See https://github.com/CLARIAH/awesome-humanities-ontologies
[10] See https://www.cidoc-crm.org
[11] See https://github.com/theme-ontology/theming

The ontology contains 6,318 concepts in total. The structure begins with 20 high-level concepts, and each one branches into multiple subconcepts. These branch into further subconcepts, and so on, to three or four levels deep. The high-level concepts can be considered broad categorisations, and the lower levels are more fine grained. For example, the high-level concept [*distribution*] branches into 24 subconcepts, including [*market*], [*theatrical-window*], [*audience-development-strategies*], and [*on-demand-screenings*].[12] Similarly, the concept [*market*] branches into nine further subconcepts, such as [*digital-market*], [*emerging-market*], and [*exclusive-content*]. In addition to this hierarchy of concepts (vertical relationships), the ontology also has concepts that are related across the different hierarchies (horizontal relationships). For example, [*digital-market*], which is a concept within [*distribution*], is related to [*developing-audiences-is-hard-to-measure*], which is a concept within [*audiences*]. It is expressed as [*developing-audiences-is-hard-to-measure*] → [*through*] → [*digital-market*] whereby the relationship between the two concepts is [*through*] and the relationship is logically in one direction only. The abstractness of many of the concepts requires labels that are less than pithy, such as the phrase [*developing-audiences-is-hard-to-measure*].

Some of the 20 high-level concepts describe physical, measurable things, such as [*distribution*], [*organisations*], [*exhibition*] and [*screens-and-media*], but the majority of the high-level concepts describe abstract things related to emotions, ideas, contexts and experiences, such as [*viewing-practices*], [*reasons-for-watching*], [*experience-and-memory*], [*social-practices*], and [*emotions-from-film-watching*]. The latter concept, for example, branches into 63 subconcepts related to specific emotions that might be experienced when watching a film, such as [*enjoyment*], [*connection-with-character*], [*inspired*], [*nostalgia*], [*confusion*], [*embarrassed*], [*wanderlust*] and [*connection-with-place*]. And again, the latter concept branches further into 12 subconcepts related to feelings about place, such as [*imaginative-travel*], [*memory-of-place-holiday*] and [*felt-like-being-there*]. Even the high-level concept [*film*], which branches into comprehensive taxonomies of film titles and genres, also has abstract subconcepts. For example, the [*film*] subconcept [*attributes*] branches into 344 descriptions of film style, theme and mood; and the subconcept [*film-interpretations*] branches into 14 descriptions of interpretative approaches to the meaning of film.

The film audience ontology was used to code datasets of varied, primary and secondary source data:

- 200 semi-structured interviews with members of the public who self-identified as people who watch film in order to explore their personal film journeys;

[12] The format in which ontology concepts are presented in this paper, using [....], is not intrinsic to ontologies but used to aid the reader.

- 27 semi-structured interviews with film industry and policy experts in order to examine the challenges and concerns of people working within specialised film exhibition, distribution and investment;

- 16 film-elicitation groups[13] in order to examine the narratives and thematic components of specialised films and to better understand the interpretative resources that audiences bring to watching film;

- a three-wave survey with a total of 5,935 responses in order to capture patterns of film watching over time;

- and an analysis of 114 key film industry and policy documents in order to identify sectoral trends around specialised films.

Having acquired these datasets we produced transcripts and plain text versions of the qualitative, long-form data (the interviews, film-elicitation groups and industry documents), and then coded the contents of these using concepts within the ontology. Here, *coding* means labelling words, phrases and sentences within the data with concepts that appear within the ontology. For example, if an interviewee talks about how some films have changed their opinions, this part of the dialogue will be coded with the concept [*film-can-change-opinions*]. The three-wave survey was nominal data and structured so that its questions and answers would automatically map to concepts within the ontology.

After the datasets had been coded using the ontology we were able to extract the coded content into a single database in order to facilitate structured searching of the sources, faceting (extracting shared qualities from the data), statistical analysis and data visualisation.[14] By extracting coded content into a database whose structure mirrored the concepts within the ontology we were able to utilise the affordances of databases – highly structured data which can be interrogated quickly – whilst always leading the user back to the full text of the sources in order to contextualise the content. For example, a user is able to retrieve 302 instances in which people have talked about [*learning-through-film*], across three datasets (audience interviews, expert interviews and film-elicitation groups). Also, the highly structured nature of the ontology means that the user can retrieve data based on conceptual relationships in addition to conceptual coding. For example, an expression made by Petunia during her interview[15] might be coded as [*learning-another-language-through-film*], but the ontology structure means that a query for [*experience-and-memory*] will also retrieve Petunia's interview, even though it is never coded with this concept, because the concept [*learning-another-language-through-film*] is a subconcept of the high-level concept

[13] The film elicitation technique involves showing selected film clips to a group of people and then asking questions to establish their interpretive responses to the clips. See Forrest, 2023.

[14] For examples, see https://www.beyondthemultiplex.org/about/visualise-the-data

[15] See https://www.beyondthemultiplex.org/view/interview?idkey=SW_HR_25

[*experience-and-memory*]. Here the ontology is serving as a metalayer that gives the computer an understanding of what is *meaningful* in our sources, and how these meaningful words, phrases or sentences are related to other concepts.

Why use a data ontology to study film audiences?

The film audience ontology could be considered to be simply an exercise in text classification using multiple, related taxonomies. However, whereas taxonomies are typically lists of terms in which the hierarchies are implied but unnamed vertical relationships, a data ontology attempts to go beyond this by explicitly naming the relationships between concepts, vertically but also horizontally, and the rules that govern them, in order to model the knowledge domain at a meta level. Therefore an ontology should be able to tell us how film and film audiences interact from the model alone, without us having to apply the ontology to any real data, just as we can understand how family relationships work from my kinship model above. Taxonomies cannot do this. If we conceive of taxonomies as collections of nouns and adjectives, an ontology is what happens when we introduce verbs and prepositions – we are able to understand how the nouns and adjectives relate to one another. For a knowledge domain, this means we are able to describe how things that exist in the real world relate to one another using a model that the computer can interpret. Developing a data ontology is a process of worldbuilding for the benefit of the computer, in order to help overcome a problem which computers experience in the real world: the semantic gap.

The challenge facing most machine-assisted research in the arts, humanities and social sciences is the semantic gap in socio-cultural data. A computer can trawl vast amounts of information, quickly and with great precision, but it does not know what the information means in the sense of how it signifies a larger, real world. In other words, when searching for 'hollywood' in a collection of texts the computer is searching for a string of nine characters; it does not know that it is searching for a place or a film industry. As humans we know what 'hollywood' means, and we have a deep understanding of its nuances and complexities with respect to the real world which it is referencing. However, unlike the computer, we lack the speed and precision to apply this understanding across enormous swathes of information. This is the semantic gap: how can we enhance a computer's speed and precision with our knowledge and deep understanding of the real world when analysing large amounts of cultural data, so that it can help us to answer complex research questions quickly, accurately, and meaningfully? Research questions such as: how are an individual's changing interests, consumption patterns and sharing practices in film and cinema shaped by their life experiences?

When working with socio-cultural data the semantic gap frequently manifests as three practical problems: how can a computer understand what data means; how can a computer reach this understanding across different types of data; and how can a computer do so if not all the data is present?

1. How can a computer understand what data means?

The majority of arts, humanities and social science research that involves the use of computational methods and digital tools uses some form of annotation or re-structuring of socio-cultural data in order to render it more meaningful to a computer. For example, tagging the text of a film script using the TEI XML markup language[16], summarising a programme of regional investments in local cinema using a database, coding an interview transcript using NVivo[17], or describing a film clip using folksonomic keywords. These metadata approaches aim to identify data that is considered significant within the knowledge domain (as opposed to other, surrounding data) and to convey what the data *is* using classificatory labels. This immediately extends the semantic, interpretive capabilities of a computer. If every reference to a film is labelled *film*, a search engine (for example) can retrieve films called *Vertigo*, as distinct from the soundtrack, a condition of dizziness, or the wordless novel by Lynd Ward; but it can also retrieve references to all films, irrespective of their titles.

The drawback with annotating socio-cultural data in this way alone is that the results are semantically quite limited beyond basic discoverability. A search engine that can retrieve one thousand documents containing references to a film called *Vertigo* does not get us very far. The computer knows that there is a film called *Vertigo*, and it knows that there are other films, but it does not know anything else about *Vertigo* (its plot, actors, director, genre, reception, cultural significance *etc*) and it does not know how any of these films are related. You might ask why this matters, because as end-users we can read about *Vertigo* by consulting one of the sources that the search engine has returned, such as an article. But if we want to start asking complex and interesting questions about the films of Alfred Hitchcock and their influence on cinema, for example, we need a computer to understand *Vertigo*'s relationship to cinema at scale, across many thousands of films, television programmes, books, articles, and interviews. This requires us to go beyond simple classification and instead code significant concepts in terms of their characteristics, relationships to one another (beyond simple hierarchisation), and rules of use (a film is not a condition that affects your balance, a film cannot be made before the director has been born *etc*). A data ontology seeks to model this complexity so that it can be used to code and add a metalayer of meaning to concepts that are referenced within our datasets. Without encoding this complexity, classification is as restrictive as trying to understand evolution simply by tagging the names of different flora and fauna.

[16] See https://tei-c.org/guidelines/p5

[17] See https://lumivero.com. NVivo was formerly a product of QSR International, now owned by Lumivero.

2. How can a computer understand different types of data?

Secondly, the semantic gap in socio-cultural data is exacerbated by the format and content of different types of primary sources, and the different approaches to digitisation which these formats need. For example, Beyond the Multiplex was a research project that required different types of primary sources to be assembled in order to capture the different dimensions of the research domain. The datasets that were acquired by Beyond the Multiplex included online survey data. Survey data is born digital and form based, in the sense that data is collected by inviting respondents to input their responses into an online form. Its native digital format is a spreadsheet (specifically, CSV) in which columns are questions and every row curtains the answers by each surveyed respondent. As such, the data is considered to be *nominal*, in that it is already labelled, already structured. By way of contrast, the interviews and focus groups were audio recorded, transcribed and then coded using NVivo. Their native digital format is an unstructured 'plain text' file (usually denoted by the file extension TXT). The interview and focus group transcripts are considered to be long-form qualitative data. Traditionally the computational task of comparing these different types of datasets is difficult. As a result, most online resources which present mixed data maintain the separation within their navigation and search, or offer federated searching by simple keyword.

However, even where the format of the primary source is the same, such as transcripts of interviews with people who go to the cinema, transcripts of interviews with people who work in the film industry, and transcripts of focus groups with people who reflect on the meaning of film, the concepts referenced within them can be quite different. For example, audience interviews will contain concepts relating to film taste, experience and memory, whereas interviews with people who work in the film industry will contain concepts relating to policy, exhibition and programming. These mixed sources and methods inhibit connectivity across the data because there is nothing which explains to the computer that people watch films that have been selected by a programmer, or that a person's childhood experience of film might influence when, where and how they watch film today. Again, a data ontology seeks to address this. An ontology serves as a metalayer that describes concepts, their qualities and their relationships separately from the individual types and formats of our datasets. So it enables us to code concepts that are being referenced within each type of dataset in the same way.

3. How can a computer understand the world when not all the data is present?

The third problem with the semantic gap in socio-cultural data is the absence of data. A computer can only form its understanding of a knowledge domain using the data that is available to it. For example, a computer cannot deduce that a person called Agatha is likely to be a woman in the absence of gender information. Similarly, if John is Agatha's brother and Agatha has a spouse called David, a computer cannot deduce that John possibly knows who David is if there is no data describing a relationship between the two. Similarly, if Agatha lives in Hull and watches films at a cinema, and David lives in Hull, if there is no data describing a relationship between *cinema*, *Hull*, *Agatha* and *David*, the computer cannot deduce that David probably goes to the cinema to watch films – it cannot infer that this is probable because David is Agatha's spouse and there is a cinema in Hull.

This ability to infer facts and relationships that are not present in the source is easy for humans and fundamental to arts, humanities and social science research where the absence of information can hold considerable significance (for example, when decolonising historical narratives). Computational pattern recognition techniques such as natural language processing and machine learning solve this problem by using large amounts of past data to inform a computer's understanding of present data – the computer learns from examples. A data ontology, on the other hand, is a model of a knowledge domain that exists separately from the datasets we apply it to, so an individual dataset might only contain concepts that appear in one part of the knowledge domain, and we use the ontology to code them as such, but when interpreting the coded dataset the computer benefits from knowing the entire knowledge domain because it has access to the entire model.

For example, Beyond the Multiplex conducted interviews with film audience members who commented that going to the cinema is reminiscent of childhood. Other interviewees believe that cinema is a middle-class place. These are two very different statements, coded with the concepts [*going-to-the-cinema-is-reminiscent-of-childhood*] and [*cinema-is-a-middle-class-place*] respectively.[18] They appear unrelated, and neither group of interviewees refer to the other concept – these are distinct beliefs. However, both are subconcepts of the concept [*cinematic-experience-in-general*], so when analysing the opinions of one group of people (those that consider cinema to be a middle-class experience) we know that other people associate the experience of cinema with childhood – of being taken there by their parents or relatives. The computer knows that the two concepts, although very different, are related as cinematic experiences, because it has access to the entire model. In fact a close reading of the extracts substantiates that there is a relationship present between the two. Being taken to the cinema as a child had no financial or class implications from the subject's perspective, because their parents or relatives were paying, whereas in adulthood they access films in other ways in order to reduce the cost. People who do not go to the cinema, for reasons that include economics, are reminded of childhood when they think of cinema because that is perhaps the last time when they regularly attended. They do not say that they are not middle class (in economic terms), but other parts of the ontology suggest that this could be a probable reason.

In practice, the ontology creates connections across data that are interpretable by the computer, enabling the computer to present us, the researchers, with data relationships

[18] Explore the ontology here https://www.beyondthemultiplex.org/view/entity?idkey=EXPERIENCE_AND_MEMORY by clicking on *[experience-and-memories-of-film-at-the-cinema]* and then *[the-cinematic-experience-in-general]*.

that we might not have seen had the data not been coded in this way. The computer is not doing any actual *thinking* here – it is simply querying the coding and presenting us with the results – leaving us to apply established methods of analysis to understand *why* these relationships exist.

Towards a theoretically informed, empirical model of film audiences

Data ontologies are typically models that already exist (they have been created by domain experts) and they are applied to source data. CIDOC CRM, for example, can be downloaded from the official website, added to an ontology editor (software specifically designed for creating content in line with an established ontology, such as Protégé[19]), and used to create content. For example, the project Seafaring Lives in Transition[20] explored the transition from sail to steam navigation and its effects on seafaring populations. The project downloaded CRM and used it to code concepts relating to ships and seafaring across a diverse archive of primary sources. The resulting data was then used to develop a range of search and data visualisation tools to support the research. This relationship between data ontology and source data is one in which the ontology already exists at the outset, and so it has a powerful influence on how the source data is coded. But clearly, when inventing data ontologies, the relationship between ontology and source data is less straightforward. An ontology has to model a knowledge domain, and the knowledge domain is represented by both the source data and the domain expert who is creating the ontology. So at the outset it is the source data and the domain expert that have a powerful influence over the content, shape and size of the ontology. For example, I invite you to invent a data ontology for film audiences now – using pen and paper – without reference to any sources. Your ontology is unlikely to be very sophisticated, and you will quickly realise that it is far from comprehensive. The design of a data ontology is an iterative process in which the domain expert is guided by the data, even if he or she begins with a conceptual framework, rather than by the domain expert in isolation of the data. This was probably also true when CIDOC CRM was being developed – it would have required lots of content at the outset to inform the model.

On Beyond the Multiplex, our film audience ontology was designed through the process of coding the source data. Coding the source data informed the design of the ontology. It was necessarily an iterative process, heavily influenced at the outset by what we were finding to be significant in the data – an approach that is generally referred to as *grounded theory*. This resulted in a very large, extremely detailed ontology. With hindsight, this was bound to be the case, since we were designing a data ontology to describe film audiences and we were reading hundreds of film audience interviews. Further, we were not designing an ontology that just described physical, measurable things in the real world, such

[19] See https://protege.stanford.edu
[20] See https://www.sealitproject.eu

as films, cinemas, and people, but also ideas, emotions, and memory which have far more variety, nuances and are harder to classify. It resulted in some concepts which might be considered duplicates, such as [*shared-experiences-of-film*] → [*associated-with*] → [*watching-film-with-partner*] and [*shared-experiences-of-film*] → [*if*] → [*watching-film-with-partner*]. Offline, the project focussed, reduced, harmonised and refined the ontology's 6,318 concepts using thematic analysis in a process that is sometimes referred to as *closed coding* as distinct from the open coding that the ontology represents (you can see the results in Wessels et al., 2023, pp. 171). Any future iteration of the data ontology would instantiate this offline process.

We applied the ontology to our source data using NVivo, which is an unusual choice of software. One would typically use an ontology editor or an XML editor. However, an ontology editor is predicated on an ontology already existing before source data is coded, and ontology editing software requires ontologies to be written as a machine-readable schema using an ontology language, such as OWL (Web Ontology Language[21]). This makes the process of iterative design-through-coding slow and cumbersome, because every new concept of interest that we come across in the source data requires us to formally define it in the ontology schema before we can code it. NVivo, on the other hand, lies at the opposite extreme in its complete flexibility and absence of rules and constraints when developing 'codebooks'. Our approach was in effect to iteratively develop an NVivo codebook whilst coding the source data, in line with an initial, outline coding framework, and to then transform this into a data ontology later, when both the codebook and the coded source data were exported out of NVivo for use in our data platform.[22]

This workflow raises a question: have we produced and used a data ontology or have we simply coded qualitative data using NVivo? Is our ontology different from an NVivo codebook? At one level we have simply created and used a codebook within NVivo using a grounded theory approach, and the size of the codebook suffers from some repetition and inconsistencies in parts, as mentioned above – problems that would be less prevalent if using a formal ontology language. The project mitigated some of this by using traditional validation and reliability research methods such as researchers comparing each other's coding and participant validation workshops when analysing the data. However, it is when the source data and codes are exported from NVivo and used within our data platform that the coding becomes a data ontology, because we use the codes in ways that typify an ontology but which are not possible in NVivo, or indeed, in any other form of NVivo codebook reuse. For example, we are able to build, analyse and then interpret a complex relational model of film audiences, with named vertical and horizontal relationships between concepts, without the coded source data needing to be present. An NVivo code

[21] See https://www.w3.org/OWL/

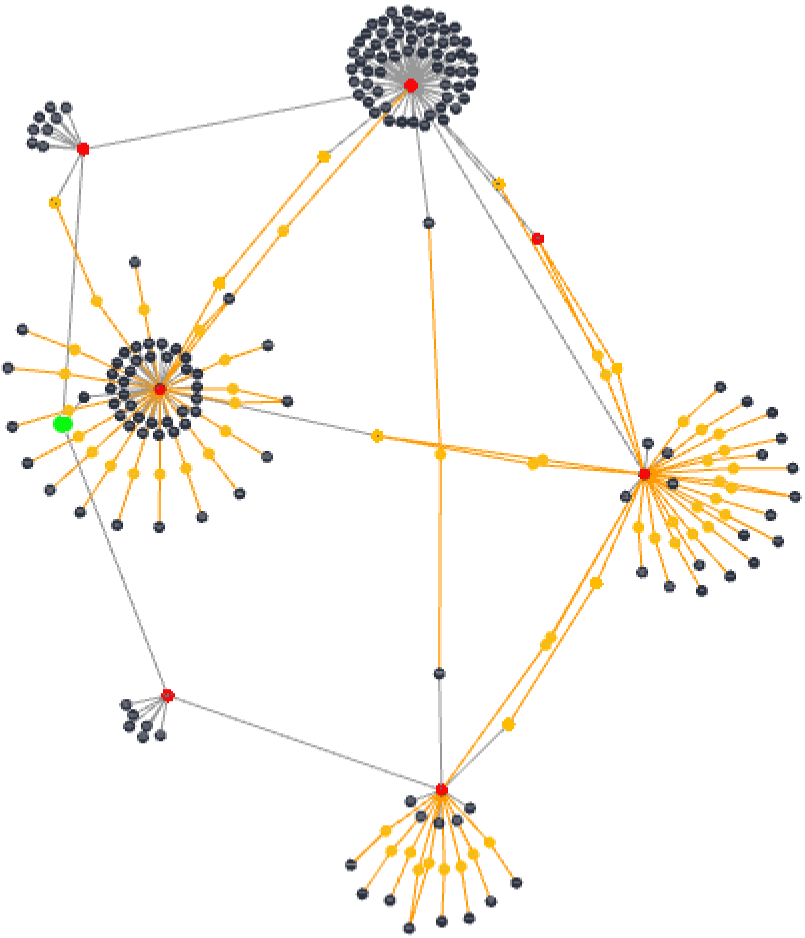
[22] This was not without its problems, because NVivo does not currently allow data and codes to be exported as a single combined file. Instead, we were required to export the full text of the source data and the NVivo nodes (lists of coded content) as two separate file collections, and then programmatically combine the two to create a single file collection.

book cannot offer this, in its 'codebook' form, because it is just a series of unrelated *nodes* (to use NVivo parlance).

Here are some example nodes from the NVivo codebook for the two related concepts [*learning-about-own-culture-through-film*] ⟶ [*opposite-to*] ⟶ [*escapism*]. We can see the linear hierarchy from the high-level concepts [*experience-and-memory*] and [*reason-for-watching*].

> EXPERIENCE AND MEMORY\Learning through film\Learning about own culture through film
>
> REASON FOR WATCHING\Reason for watching film [generic]\Escapism
>
> Relationships\\Learning about own culture through film (OPPOSITE TO) Escapism

Here are the same nodes transformed into a data ontology outside the NVivo system. They are represented using a network visualisation.



The green node on the left of the network is the root of the ontology (a parent concept simply called [*ontology*]). The red nodes are the concepts for [*learning-about-own-culture-*

through-film] → [*opposite-to*] → [*escapism*] which we can trace around the network. The black nodes are other types of concepts and the orange nodes are other types of relationships. The highly relational nature of the ontology is immediately evident from this visualisation when compared to the flat, linear hierarchies of NVivo.

However, our data ontology is still powerfully influenced by its origin as an NVivo codebook, and by an approach to ontology creation that is based on grounded theory. Grounded theory, in which theory is developed based on close readings of the data, necessarily results in an ontology that models an emerging understanding of film audiences. And given that we set out to code ideas, emotions, experiences and memories, our process of interpreting the data is more likely to be reflected in the ontology because of the semantic ambiguities that can occur within the scope of this knowledge domain, unlike the coding of physical, measurable things. As already mentioned, the project's offline work in thematic analysis could be instantiated in a future version of the ontology so that it becomes a theoretically informed, empirical model of film audiences, rather than its current status as a prototype model of film audiences. Of course, this is perhaps a state of being that all data ontologies experience, when they are being created, until they have been reused, reduced and refined.

Did the ontology tell us anything?

Clearly, annotating a large dataset of primary sources will enable us to undertake macroscopic analyses and to discover things that cannot be seen by reading the data sequentially, word by word – insights that Franco Moretti coined as *distant reading* (see Moretti, 2000). These insights are often patterns, trends and anomalies that are then best communicated using data visualisations. It also enables us to discover things that we would always have found through sequential, close reading, but we accomplish the discovery at greater speed. And of course, annotating a large dataset of primary sources will mean that we miss things – the things that were not annotated because the grounded theory approach did not recognise them as significant – which is why distant reading should always be supported by close reading and reflexivity of our methods. So did our use of a data ontology on Beyond the Multiplex produce research findings that we might otherwise not have found? For example, could we have chosen to simply read the dataset, take copious notes, perform keyword searches, or annotate the data using a taxonomy of concepts that describe only physical, measurable things, and still have produced the same results? There are many other ways to annotate a large, varied dataset too, such as an NVivo codebook, folksonomies, or computational pattern recognition techniques such as named-entity recognition and machine learning. Did our data ontology produce research findings that would not have been discovered if we had annotated the data using these other approaches? The answer to all of the above is yes, we did learn things that would have been unlikely or harder to be discovered had we not annotated the data, and it was specifically our use of a data ontology, as distinct from other types of annotation, that led us to our findings.

The principle research finding of Beyond the Multiplex constitutes a paradigm shift in how film audiences are understood. Whereas previous film audience research has always studied audiences as a group of people that already exist (*X* number of people with *Y* characteristics who are watching a film called *Z* in a specific location), our project showed that it is possible to study how audiences *form*, in the beginning, as a process, and that this process then informs our understanding of what the audience is as a group, its shared characteristics and motivations. For example, we were able to challenge the prevailing view that specialised film is only watched by educated middle-class people. If we attend to how audiences form around specialised film, the data shows that most people develop a personal relationship with different types of film through different film journeys (activities, experiences, identities, memories) and these journeys are enabled and influenced by different kinds of access to film, screen locations, and local programming and exhibition. In short, the actions of local cinema have a direct impact on how audiences form, not by simply putting bums on seats, but by influencing our relationships and journeys with film at different life stages. As a result, people's relationship to specialised film, across all demographics, can be influenced by its provision at different life stages. This is a research finding that has policy implications, and is discussed at length in Wessels 2023 and Wessels et al., 2023.

But how did the data ontology enable us to learn this? The ontology's inclusion of emotional and experiential concepts meant that we were able to code these systematically using a conceptual structure that also enabled us to trace their relationships to physical, measurable things such as place, people, films, exhibition and programming.

For example, here is a quote from an interview with an audience member called Lenny:

> It's like a short-lived antidepressant. You go into that film and you could be miserable and a bit downbeat and you could watch a film, it could be any film, and nine times out of - nine and a half times out of ten, I could leave the cinema feeling refreshed and enjoyed. I could just feel much better.[23]

This statement was coded with the concepts [*therapeutic*] and [*good-for-you-mental-health*]. These are both subconcepts of [*reasons-for-watching*] within our ontology.

Lenny then says:

> That's one of the reasons why I just like going to a cinema 'cause for me it's just a really good escape. I can just zone out … and not care about anybody.[24]

[23] See https://www.beyondthemultiplex.org/view/interview?idkey=SW_HR_06
[24] Ibid.

This statement was coded with the concept [*cinema-is-immersive*] as a subconcept of [*experience-and-memory*]. The concept is also related to the concept [*not part of an audience*] which is a subconcept of [*audiences*]. Whereas one might expect that the experience of going to a cinema would principally be about the shared experience that comes from being part of a larger audience, it is actually the immersive nature of cinema that Lenny finds to be therapeutic.

Lenny goes on to say:

> It's not like being at home. At home, you have your own comforts, so, you know, you can go to the bathroom, you can pause it, you can go and get snacks … [But] the cinema is a big screen, it's dark, it's comfortable.[25]

Lenny's reference to a big screen is coded with two concepts: [*big-screen*] and [*immersive*]. Within our ontology, the concept [*big-screen*] is related to the concept [*immersive*] like this: [*big-screen*] → [*is*] → [*immersive*]. Here, the relationship [*is*] connects two concepts that descend from different high-level concepts: [*big-screen*] is a subconcept of [*screen-and-media*] while [*immersive*] is a subconcept of both [*film*] and [*experience-and-memory*]. So the relationship [*is*] connects horizontally between things that are experiential and things that are physical and measurable. This matters because a traditional annotation model might code both types of concepts, but the model would not make the relationship between the two explicit, in a way that is interpretable to a computer. Further, Lenny never uses the word *immersive* in his interview, but this is the meaning we interpret from his explanation of the big screen, and so we have coded it as such. With our data ontology a computer is able to infer that people such as Lenny are part of an audience that gains value from the isolating, immersive qualities of the darkness and big screen, even though this is a shared, communal experience.

If taken at face value, we have a set of coded concepts here that might be familiar to anyone who has developed and used an NVivo codebook. But what is specifically 'ontological' about the coding is that the concepts form part of a larger model that has been developed through the process of coding many other interviews. The model shows Lenny's motivations and experiences are shared by others, but not universally. Other interviewees have different perspectives – they feel that it is the shared experience of cinema that is therapeutic. It shows that a person sitting next to Lenny in the cinema might be enjoying the experience because there are people like Lenny around them, whereas Lenny is zoned out and oblivious to those sitting next to him. Importantly, the structure of the ontology enables us to relate thousands of experiences to high-level concepts such as [*audiences*], [*viewing-practices*] and [*experience-and-memory*], but using subconcepts that give insights that are much more fine grained. So even though Lenny never explicitly states that he is part

[25] Ibid.

of an audience – in fact, he sees himself as being alone in terms of his actual experience – the ontology infers that he is part of an audience from the relationship of concepts within his interview to concepts within other interviews – they share the same high-level concepts. Further, we can see that for people like Lenny, the audience experience – of immersive isolation – is in part due to a horizontal relationship with concepts that denote physical, measurable things, such as big screens and darkness.

Our data ontology shows us that audiences exist, but that they are made up of groups of people with shared characteristics. Audiences are diverse – we knew that – but we can see how and why they are diverse. This has resulted in the project's typology of five audiences, as described in Wessels et al., 2023, pp.174-175. If we simply read all the interviews we would never deduce this; and if we restricted ourselves to an NVivo codebook we would never understand how the concepts are related as an overall model of the knowledge domain. We would never understand, for example, that it is the immersive quality of cinema that makes some people feel that they are not part of an audience when they watch a film, and that this is a shared audience experience that exists alongside other audience identities simultaneously, influenced by the physical provisions of cinema.

Conclusion

Developing a data ontology to study how audiences form was an ambitious component of Beyond the Multiplex's research methodology, and potentially high risk given that the data platform and a lot of subsequent data analyses relied on it. The process of developing and applying the ontology was not without its problems: the choice of NVivo as a tool for iteratively developing the model and applying it to data meant that there were some inconsistencies in the naming and organisation of concepts, as well as the occasional repetition, whilst the size of the ontology (6,318 concepts) meant that not all our data could be coded comprehensively due to time limitations within the project. The ontology suffers from a level of detail that would not be present had we used a formal ontology language such as OWL and an ontology editor such as Protégé, but the detail is also the ontology's richness. These are all valuable lessons, and the problems were mitigated using more traditional data analysis techniques. There is clearly a need to refine the ontology using the project's subsequent analyses before we document and publish it for reuse so that other researchers can use it in their own projects, but this is perhaps a creative process that all data ontologies go through. Our surprise was that we set out to establish an ontology of film audiences which we could then apply to our primary source data, but instead the data underpinned an emerging, theoretically informed, empirical model of film audiences. The ontology in its current form arises from a coded analysis of the data, rather than being a model that imposes itself on the data. In this sense, the ontology itself is a valuable research output, even if it would benefit from further work. Despite all this, the simple truth about data ontologies – that they help us to bridge the semantic gap so that computers can interpret rich, varied sources in ways that are meaningful for us – enabled us to achieve insights that would not have been possible had we used other annotation methods. Key to this was our ambition to develop an ontology that modelled the human world of ideas, emotions, memory and experience, in addition to physical, measurable things, so that we could understand how the physical and the experiential interact in the formation of film audiences. This is a domain that data ontologies rarely venture into because, as Beyond the Multiplex shows, it is a level of ambition that some might consider heroic and others foolhardy.

Biographical Note

Michael Pidd is Director of the Digital Humanities Institute at the University of Sheffield. He has nearly 30 years of experience in developing, managing and delivering collaborative research projects and technology R&D in the arts, humanities and cultural heritage domains. He has been the principal investigator, co-investigator and technical lead on UKRI projects, as well as supporting the technical delivery of over 120 collaborative projects and 70 online research resources and information systems. Michael was a co-investigator on the AHRC funded Beyond the Multiplex project which makes use of a data ontology to help in understanding film audiences.

References

Forrest, D. (2023). 'I'm no expert, but…': Everyday Textual Analysis with Film Audiences in the English Regions. *Participations*, 9(2). Available online at: <https://www.participations.org/19-02-13-forrest.pdf>.

Moretti, F. (2000). Conjectures on New World Literature. *New Left Review*, 1 January/February. Available online at <https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature>

Pidd, M. (2021). *A Practical Guide to Using Data Ontologies in the Arts, Humanities and Social Sciences*. The Digital Humanities Institute, University of Sheffield. Available online at <https://www.dhi.ac.uk/books/ontology-guide>.

Wessels, B. (2023). How Audiences Form: theorising audiences through how they develop relationships with film. *Participations*, 9(2). Available online at: <https://www.participations.org/19-02-16-wessels.pdf>.

Wessels, B. Merrington, P., Hanchard, M and Forrest, D. (2023). *Film Audiences: Personal Journeys With Film*. Manchester: Manchester University Press.