THE ORGANISATIONAL IMPACT OF AN ARCHIVE OF NEWS SUBTITLES: USEFULNESS AND ACCESSIBILITY

CARRIE HICKS

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Information and Library Studies

DEPT. OF COMPUTER AND INFORMATION SCIENCES

UNIVERSITY OF STRATHCLYDE

AUGUST 2019

DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself. Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.

(please tick) Yes [ ] No [×]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices) is 21,995.

I confirm that I wish this to be assessed as a Type 1 [✓ 2] 3 4 5 Dissertation (please circle)

Signature: Carrie Hicks

Date: 18/08/2019

Abstract

In acknowledgement of the fact that broadcast archive footage can be difficult to access for external users, this exploratory research considers the possibility of utilising subtitle files as an alternative means of obtaining relevant content. This research takes place within the context of STV news, a Scottish broadcasting company which has a wealth of historic subtitle files that are preserved but largely untouched in an unused archive.

The aim of this research was to discover what steps need to be taken in order to transform subtitles from a secondary support medium of communication into a useful resource in their own right. This involved the creation of descriptive metadata to make the content contained within the subtitle files discoverable. By doing this, the intention was to identify problem areas and opportunities for improvement, from which future research could draw upon.

It was discovered that subtitles have potential value as a source of proxy content. However, the characteristics of subtitles, as a textual medium that also represent audio content, brings together a unique combination of features. The metadata originally created needs further refinements before an accessible subtitle archive could be successfully established and maintained.

Acknowledgements

I would like to thank my supervisor, Dr Martin Halvey, for all his advice and guidance throughout this project.

List of illustrations

Contents

1.0. Introduction

Large-scale operations are increasingly underway to digitise analogue material in order to safeguard it from obsolescence (Schuller, 2015). Digitisation is said to offer great potential for accessibility, as users are not restricted by the analogue format of resources, nor limited by their physical proximity to said resources (Balbi, 2011). However, this promise is not always fulfilled for reasons related to how the material is organised. For further discussion on this, see sections 2.1.2 - 2.1.4. Within the broadcasting industry, there are further obstacles to accessibility. There is no requirement for these institutions to provide external access to their collections, nor are they set up to accommodate such requests (Kramp, 2014). As a result, broadcast archives remain relatively closed off. Consequently, extensive collections of audiovisual material, as well as historic written ephemera, are largely inaccessible to external users (Hewett, 2014).

The closed captioning script that accompanies television footage could offer a potential solution to this problem of inaccessibility. More commonly known as subtitles in the UK, their primary purpose is to serve deaf and hard of hearing communities by providing them with a textual representation of the audio content. Therefore, subtitles act as a substitute for the broadcast footage. However, it seems this potential solution is not being utilised. While several authors do rely on proxy content to access television footage, they tend to use either abstracts or transcripts to do so. Examples of such studies include Kuklinski and Sigelman (1992) and Maguire (2002), although for a full discussion on proxy content, see section 2.5.

There has been little mention of re-appropriating subtitles for the purpose of using them as proxy content. Instead, much of the research on subtitles up until now has focused on using them as an index for audiovisual content (Brown *et al*., 1995; Cohen and Rosenzweig, 2006; Heeren, Ordelman and de Jong, 2008). This dissertation seeks to explore what steps need to be taken to transform subtitles from a secondary support medium into a useful resource in their own right.

Apart from Snider (2000), there has been little discussion on the value of subtitles as a useful resource. However, his work focuses on the US broadcasting industry and is purely discussion-based. Although he argues that a subtitle archive would provide a valuable public service, he offers only limited guidance into how such an archive could be created. In nearly 20 years since this paper was published, there is nothing in the scholarly literature to suggest Snider's idea is any closer to being achieved. This illustrates the need for research into the practical realities of creating a subtitle archive.

1.1. Research questions

This work takes the form of a case study of STV news. STV is a Scottish television channel which serves north and central Scotland, providing extensive regional news coverage (STV, 2017). In accordance with UK legislation, the channel is obligated to subtitle 90% of its programming in order to comply with Ofcom's code on television access services (Ofcom, 2017a). Since 2016, I have worked at STV as a subtitler, responsible for providing accurate textual representations of the company's broadcast output. Once aired, the subtitle files are stored in a "dark archive" (Allen and Johnson, 2008, p. 394). That is, they are preserved but untouched, with no easy or convenient way to access to the content. Ubois (2006) argues that for content to be considered fully accessible, it must be discoverable. In order to make the content discoverable, descriptive metadata must be applied (Zeng and Qing 2008). For a fuller discussion on discoverability, see section 2.7.

It is my experience of working as a subtitler that has driven this research. I am interested in improving access to STV news subtitle files by creating metadata that will enhance their discoverability. By doing so, I hypothesise that a subtitle archive could have some practical value outside of its original purpose as an accessibility service for the hearing impaired. In this dissertation, I seek to explore the benefits and challenges of setting up and maintaining an archive of news subtitles. This research will address the following questions (RQ):

- RQ1: What is the value of an archive of television news subtitles?
- RQ2: How can we enhance discoverability of content in a subtitle archive?
- RQ3: To what extent can we retroactively annotate records from the past to provide added value?
- RQ4: What is the trade-off of implementing an accessible subtitle archive in a real-life context?

1.2. Research objectives and outcomes

The overall goal for this dissertation is to develop an understanding of the initial set-up of a subtitle archive, and the associated challenges that arise during this preliminary stage. It is hoped that this could serve as a starting point from which to develop a comprehensive and fully realised version of the archive.

To answer the research questions, the following objectives (Ob) are put forward:

- Ob1: Explore the barriers faced by users who want to access archived news media content
- Ob2: Create metadata for news subtitles in a way that is compatible with how it would be carried out in a professional context
- Ob3: Compare and contrast historic and current news samples that are broadly comparable in terms of content

I will focus on Ob1 throughout chapter 2 of this dissertation, which is devoted to building the context in which RQ1 can be addressed. The findings of chapter 2 will be synthesised in section 2.9. Ob2 sets out to answer both RQ2 and RQ4, while Ob3 will address RQ3. RQs 1-3 will be addressed in the analysis chapter, where I identify four main challenges that arise from the metadata creation process, and put forward some suggestions for improvement. I will refer to RQs 1-3 in the analysis and suggestions subsections of chapter 4. In the conclusion, I will re-address all the research questions and provide some thoughts on future directions for research.

2.0. Literature review

The purpose of this chapter is to review the literature on accessibility of news content. I begin by providing an overview of print media, and the factors which impact on accessibility of textual news content, in both analogue and digital formats. This is relevant to my research because subtitles, like traditional print, are textual in nature. Yet subtitles also differ from traditional print media as they are representative of spoken content. Therefore, the remaining sections will focus on news in an audiovisual context. I will draw parallels with both forms of news media, print and television, throughout this dissertation.

2.1. Traditional print

2.1.1. Historical print collections

Newspapers have a long history of archiving their print collections. In the United Kingdom, legal deposit legislation compelled newspaper publishers to provide libraries with copies of their output (Fleming and King, 2009). As a result, comprehensive collections of regional, national and overseas newspapers are available today. Currently, the British Library's newspaper archive holds over 600 print volumes, dating from 1603 to the present day (A unique archive, 2019).

Although these collections offer great potential for historic research, access is constrained by the print format. In order to use the collections, users must travel to the premises which hold the content. In addition, the manual search process is extremely laborious, and can be further prolonger by the absence of a useful system of organisation. Hansen and Paul (2015) conducted a study into the archiving practices of nine news organisations. Although most preserved a full run of their print newspapers, they often had no accompanying index to the contents (Hansen and Paul, 2015). Thus, preservation of content does not always entail access to the content.

2.1.2. Digitisation to improve accessibility

The manual search process of analogue material limits accessibility, thus providing a compelling reason for digitisation. In recent years, there has been a rapid increase in digitisation initiatives around the world (King, 2005; McKernan, 2014). Projects such as these not only help preserve the original copy but also widen access to the newspaper collections. It has been argued that digitisation supports a more democratic approach to knowledge as patrons are no longer limited by physical accessibility and geography (Balbi, 2011). However, digitisation brings its own limitations in terms of accessibility, which I will discuss in sections 2.1.2.1 and 2.1.2.2 below.

2.1.2.1. Keyword searching

Textual content can be converted into a searchable index (Mussell, 2017), which has advantages in its simplicity and convenience (Tanakovic, Krtalic and Lacovic, 2014). However, a limitation of this is that some useful resources may be overlooked if the searcher enters a term that is not used in the document. As Bingham (2010) observes, the absence of a keyword or phrase does not necessarily indicate that the subject has not been referred to. Therefore, other forms of metadata are needed in order to help users locate relevant content (Mussell, 2017).

2.1.2.2. Collaborative tagging

One method of adding metadata to content is through collaborative tagging, in which users assign their own keywords to documents. This allows others to search and browse the content based on these terms (Macgregor and McCulloch, 2006). The British Newspaper Archive encourages users to tag articles in this way (What are tags for, 2019). However, there are conflicting findings on the effectiveness of collaborative tagging. Holley (2010) evaluated the use of collaborative tagging on full-text searchable historic newspapers. Over 500 users took part in the process, and 38,000 distinct tags were created. She found the feature was especially well-received by serious researchers. The tagging was not moderated by staff and thus, offered a cost-effective method of providing value added content. Despite the lack of moderation, no abuse of the tagging feature was found. In contrast, Razikin *et al*. (2011) found that unfamiliarity with the system resulted in low quality user-created tags which did not aid others with resource discovery.

2.1.3. Vocabulary control

User-assigned and free-text keywords suffer from a lack of control over synonymy, misspellings, and variant forms of the same word. However, these can be overcome with the use of a set list of predefined subject terms, known as a controlled vocabulary (Joudrey, Taylor and Miller, 2015). Cox, Tadic and Mulder (2006) outline several established controlled vocabularies. Of these, the Library of Congress is a widely used source. It maintains controlled data values for subject headings (LCSH) and name authority files (LCNAF). However, as McCutcheon (2009) notes, LCSH are not well-served for certain subject domains. Furthermore, controlled vocabularies may not meet the needs of localised communities or organisations (Macgregor and McCulloch, 2006). For more discussion on vocabulary control, see section 2.7.1.1.

2.1.4. Discrepancies between print and digital collections

A further problem for users can arise from discrepancies between print and digital newspapers. This was first observed by Kaufman *et al*. (1993) who found inconsistencies in which newspaper edition was uploaded to the electronic database, resulting in different content between both versions. Snider and Janda (1998) note that articles may be removed from databases if they become subject to legal challenges.

Some databases are text-only. This is a factor which ought to be considered when using digital newspapers as a scholarly resource, as imagery can offer meaningful insights into journalistic practices. Maurantonio (2014) compared print and digital copies of newspapers to analyse coverage of a 1985 Philadelphia bombing. She found that the photographs in the original print version added meaningful context for the stories, which was missing from the digital counterpart. In some cases, legal restrictions may impose a lack of visuals. O'Connor (2015) analysed a repository of government publications and found that many images were redacted due to copyright infringement.

These problems do not negate the value of digital newspapers. Digitisation allows users to access, read and analyse texts in new ways. It has been argued that this process of remediation turns the digitised image into a resource in its own right (Nicholson, 2013; Mussell, 2017). Using this interpretation, digital newspapers can offer users a valuable tool for scholarly research, but must come with the understanding that they are not merely an exact copy of print newspapers in a digital format.

2.2. Audiovisual media

So far, this chapter has discussed preservation and access from the perspective of textual media. The remaining sections in this chapter will focus on audiovisual media, as this is where the role of subtitles becomes relevant. Before drawing conclusions about the value of a subtitle archive, it will first be necessary to provide context for the historical, cultural and technical reasons that have impacted on the accessibility of broadcast archives.

2.2.1. History of audiovisual preservation

Newspaper archives, both print and digital, present challenges in terms of access. However, they have long been established as a valuable resource for historical research and libraries have endeavoured to make these collections available to patrons (Massis, 2012). In contrast, access to historical audiovisual material is limited compared to print media (Lessig, 2004). Some explanation for this can be traced back to the early days of television broadcasting, when technology and attitudes shaped preservation efforts.

Initially, the high cost of film meant that material was preserved very selectively (Looking for old programmes, 2017). Whannel (2005) provides a good example of this with regards to the recording of sports coverage, noting that film cameras would only be turned on in a football match if a goal seemed likely. Consequently, unexpected action and events were not captured on film. The arrival of videotape did not increase preservation efforts, as television broadcasts were considered ephemeral and videotape was wiped for re-use (Murphy, 1997; Martin, 2005). Despite the uptake of videotape, television was afforded little respect and was not perceived to have much historical value (Spigel, 2005). News footage was also not preserved as it had no resale value (Looking for old programmes, 2017).

Considering the perceived value of television, combined with attempts to keep costs low through the re-use of videotape, it becomes evident why much historical television content has been lost. However, the introduction of policy at the British Film Institute (BFI) helped prevent further catastrophic loss (Madden, 1981), and further legislation in the 80s and 90s allowed the institute to build up a storage of off-air recordings (Bryant, 2010).

2.2.2. Obsolescence

Today, attitudes towards television are at odds with those of the past and there is a recognised need for preservation of television content (Council of Europe, 2001). There is a consensus among many authors that television has social and cultural value (Spigel, 2005; Allen and Johnson, 2008; Kramp, 2014; Assmann and Mearns, 2015). Schuller (2015) observes that, along with the radio and film industries, broadcast television showcases diversity of language and culture. The merit of television is no longer in doubt. However, there is now a threat posed by obsolescence. The term obsolescence can be applied to recording formats that are decaying due to age, or equally, products that are no longer being produced because of a lack of consumer demand (Assmann and Mearns, 2015).

Analogue recording carriers are susceptible to physical deterioration, although with proper care and storage, they can last several years. However, without their associated playback equipment, which is becoming rapidly obsolete, they are worthless (Schuller, 2015). An early example of technical obsolescence is the BBC Domesday Project, a digital resource created in 1986 about life in the UK during that time period. The storage medium was videodisc and, though the discs themselves were well preserved, the hardware and software needed to read the discs became obsolete. However, with significant effort, the authors were able to preserve the content contained on the discs (Darlington, Finney and Pearce, 2003).

This raises an important difference between information content and information carrier. The information carrier is the device on which the content is stored and transmitted (Feather, 2018). The carrier itself may also have value. However, in the case of the BBC Domesday Project, the true value came from the stories and memories contained on the discs, or, in other words, the information content.

I would argue that subtitles act as an information carrier, as they preserve the same information content found in the original broadcast footage, albeit rendered in a textual medium rather than an audiovisual medium. Therefore, an archive of news subtitles has value as an additional means of safeguarding the information content.

2.3. Approaches to preservation

In the literature on digitisation, the comprehensiveness of preservation policies has been subject to debate (Assmann, 2011). This has led to two opposite approaches - the all-in approach and the slow-play approach (Balbi, 2011). The former supports the idea that everything should be kept. Proponents of the slow-play approach advocate for a more selective method, in which only some items are preserved as a representative sample (Balbi, 2011).

2.3.1. The all-in approach

This approach argues that the role of the archivist should not be to judge the perceived merit of a resource (Amernick, 2018). In some ways, this approach is the simpler of the two options as no value judgements need to be made, thus expending less mental energy. Some authors also argue that this approach is more democratic, ensuring the widest possible knowledge is available to the public (Assmann, 2011; Kramp, 2014).

As discussed in section 2.2.2, television has value as a social and cultural artefact (Allen and Johnson, 2008; Schuller, 2015). In addition, it offers great potential as a resource for scholarly work (Murphy, 1997; Hale, Fowler and Goldstein, 2007). This provides compelling reasons for developing broad and comprehensive preservation policies.

However, digital collections are growing at an exponential rate (Feather, 2018). This comes with the associated problem of retrievability. Without an accurate index, finding relevant digital content is a laborious task (Wactlar and Christel, 2002). Thus, the all-in approach to preservation puts an unreasonable demand on archivists to process the material and make it available (Feather, 2018).

2.3.2. The slow-play approach

In contrast to the all-in approach, a smaller and more selective approach to preservation allows for more time to be spent creating high quality metadata, thereby improving its accessibility (Compton, 2007). Rooks (2010) notes how a transition from the slow-play approach to the all-in approach resulted in a trade-off metadata quality (Rooks, 2010). Thus, resources lose value if the content contained within them is not discoverable. However, a selective approach to preservation cannot anticipate potential current or future research interests (Murphy, 1997).

Although there are advantages and disadvantages to both approaches, I will follow the all-in approach with regards to my own data samples. In order to fully assess the suitability of subtitles for a proposed accessible subtitle archive, it will be necessary to examine the data in its entirety to identify problem areas and opportunities for improvement.

2.4. Accessibility of broadcast archives

Having discussed approaches to audiovisual preservation, I will now return to the issue of accessibility. In section 2.1, I described some of the factors which impact on accessibility of traditional print media. I will now consider accessibility of audiovisual material, from the perspective of the institutions which own the content. I will discuss these issues as they pertain to audiovisual archives generally, rather than focus solely on news archives. This is because the issues that arise have widespread applicability. This will provide further justification that an accessible subtitle archive could provide an alternative means of access to broadcast content.

2.4.1. Institutional concerns

Digitisation means that technical constraints need no longer be an obstacle to access (Wright, 2009). However, the reality is that audiovisual archives are not set up to accommodate external access (Kramp, 2014). The priorities of broadcasting are production and delivery. Thus, the primary aim is preservation of material for potential commercial exploitation (Cigognetti, 2001). The development of new BBC services in the late 90s saw an increase in re-use of archival material. For example, the schedule for the BBC 7 radio channel primarily consisted of archive news and drama content (Rooks, 2010). With such great demand placed on the archivists, they would have little time to assist with external requests that do not serve the needs of the institution.

2.4.2. Barriers

Access may be granted to audiovisual archives on an individual basis, although broadcast institutions are under no obligation to provide this (Kramp, 2014). In the US, broadcast archives are fragmented. Ubois (2006) and Collins (2010) both discuss having to travel to different locations to view material on-site, which comes with associated financial costs. Copyright restrictions may also limit outside access to material (Ubois, 2006; Looking for old programmes, 2017).

In the UK, the National Archives at the BFI are responsible for preserving broadcast output for ITV, Channel 4 and Channel 5 (Bryant, 2010). However, only 12.5% of the channels' content is preserved (Introduction to the BFI Collections, 2019). While the National Archives provide access to some programme content, broadcasters may, in addition, have their own in-house archives. However, this footage tends to be for business-to-business purposes, and is costly to obtain (Looking for old programmes, 2017). The ITV archive explicitly states that they will not sell footage to members of the public (ITV Archive, n.d.). Thus, options for obtaining programme content for personal, research or non-commercial purposes are limited.

2.4.3. Assumption of widespread availability

Historical broadcast content may be obtainable via DVD or online streaming services. Along with rapid improvements in digital technology, this has led to the assumption that all such content is easily and widely available (Spigel, 2005; Collins, 2010). However, the programmes and archive footage that are made available is dictated by market interests, which do not necessarily align with the interests of people who want to access archive material for non-commercial purposes (Kramp, 2014).

Many scholars have argued that only studying what is widely available can lead to researcher bias (Compton, 2007; Collins, 2010; Kramp, 2014). When trying to obtain episodes of the US programme *All in the Family*, Collins (2010) found some clips and short extracts available online. However, these tended to highlight only the show's sensationalist moments. *All in the Family* was selected for study because it served as an example of cultural change at the beginning of the 1970s. However, the available clips and extracts alone would not allow for an in-depth analysis of the programme and what it revealed about American society at the time. If scholars are restricted to evaluating commercially available content, valuable insights into television history are at risk of being neglected.

2.5. Proxy content

Having discussed the barriers that can prevent people from watching and obtaining broadcast archive footage, this section will now examine alternative methods of accessing the content. Snider (2000) proposed that a repository of news subtitles could provide a cost-effective, easily accessible substitute for audiovisual news footage. However, the research to date seems to focus on accessing news through other forms of proxy content. I will examine these content surrogates in turn before discussing subtitles in more detail in section 2.6.

2.5.1. News abstracts

The Vanderbilt News Archive was established in 1968 and features regular news broadcasts from several US networks. Patrons can stream select content, and all other material can be loaned for a fee. Each individual news item is accompanied by an abstract, which is freely available to see for anyone searching the database (About the archive, n.d.). Several studies investigating news media have used Vanderbilt's abstracts as a data source (Kuklinski and

Sigelman, 1992; Norris, 1995; Maguire, 2002; Kernell, Lamberson and Zaller, 2017). These authors fail to specify their reasons for using abstracts in their analyses, although this may be due to the difficulties of accessing broadcast footage, as discussed earlier in section 2.4. Another reason may be that it is less time-consuming to read abstracts than to watch hours of television footage.

However, there are limitations in using abstracts for scholarly research. The content is dependent on the judgement and skill of individual writers, and so may be not consistent in terms of quality (Snider and Janda, 1998). Althaus, Edy and Phalen (2002) conducted a quantitative analysis to determine whether abstracts are an effective stand-in for full-text transcripts. The results indicated that while abstracts can be used to identify broad themes and topics in news reports, they are not well suited to representing specific language use. Kernell, Lamberson and Zaller (2017) acknowledge a limitation of their study is that news abstracts may not accurately represent the contents of the original broadcast.

2.5.2. Transcripts

Whereas abstracts are summarised documents, full-text transcripts act as a closer representation of the original audiovisual footage. The term transcript is generally understood to mean a word-for-word copy of a document (McClure, 1999). Subtitles are sometimes thought of as a type of transcription but, using this definition, I would argue that this is not necessarily true. I will discuss the reasons for this in section 2.6.1.

Several studies used broadcast news transcripts in their research (Romer, Jamieson and Jamieson, 2006; Ubois, 2006; Kang, Gearhart and Bae, 2010). When Ubois (2006) was unable to access broadcast footage, he found transcripts to be the most useful alternative source. All the authors cited above used the LexisNexis database to obtain transcripts. Transcripts are available from other sources but, of these, LexisNexis was found to have the longest run of coverage (Ryan and Simon, 2014).

Despite its widespread use, LexisNexis suffers from several drawbacks. In comparison to the Vanderbilt news abstracts, the transcripts only date back to the early 90s, as opposed to Vanderbilt's 1968. A subscription is also required to access the service whereas the abstracts are freely accessible (Althaus, Edy and Phalen, 2002). Furthermore, the coverage is largely confined to US television networks, thereby restricting options for research into news coverage from other countries. In the UK context, there is no single repository for news transcripts. The BBC state that they do not produce transcripts (Can I use BBC content, 2018). I also know from personal experience that STV do not produce transcripts for news programmes.

It is not clear where the LexisNexis transcripts are sourced from, as this information is not made explicit on the website. However, one source claims that the transcripts are created by third-party providers (Ryan and Simon, 2014). It is also not clear what conditions the transcripts were produced under, and if they are truly a word-for-word copy of the audio material. This information ought to be made explicit, as this can affect how users interpret the data. For example, Romer, Jamieson and Jamieson (2006) used news transcripts to identify reports of suicides. They concluded that local TV coverage can contribute to the phenomenon of suicide contagion, in which exposure to such stories can influence suicidal behaviour in at-risk young people. A methodological weakness of the study is that the authors only searched the content for the keyword 'suicide'. As discussed in section 2.1.2.1, keyword searching can have limited value when searching textual content (Bingham, 2010). In this study, it is not known whether the transcripts used were verbatim. Therefore, it may be the case that the word count for 'suicide' did not exactly correspond with the words that were originally spoken.

2.6. Subtitles

2.6.1. Subtitles as an accessibility measure

So far, this chapter has focused on issues of preservation and access to news media, including the use of abstracts and transcripts as proxy content. Before proceeding to examine how subtitles can be re-appropriated as proxy content, it will first be necessary to discuss the primary function of subtitles as an accessibility measure for the hearing-impaired. This will help to build a picture of the issues that affect the production of subtitles, particularly with regards to news. This will highlight the differences between subtitles and verbatim transcripts. I will focus on subtitling in the context of STV in methodology section 3.3.

As discussed in the introduction to this dissertation, the term closed captioning is used as the US equivalent for UK subtitling. They are closed because they are hidden, and users can switch them on or off in accordance with their preferences. The opposite is true for open subtitles, which are permanently embedded on screen (Robson, 2004). Interlingual subtitles are used for foreign language translation. In this dissertation, the term ‘subtitles’ refers exclusively to intralingual subtitles for deaf and hard of hearing viewers. Intralingual subtitles are the textual representation of audio content. This includes spoken language, as well as paralinguistic features such as tone of voice (Neves, 2005).

The UK pioneered the development of subtitles for the deaf and hard of hearing. In the mid-70s, the BBC and ITV developed the CEEFAX and ORACLE teletext services, which enabled users to receive subtitles on their home television sets (Downey, 2008). The introduction of the 1996 Broadcasting Act and the 2003 Communications Act set high requirements for subtitling on the five UK terrestrial channels. Channel 5 must subtitle 80% of its programming, while the target for ITV and Channel 4 is 90%. The BBC has the highest requirement at 100% of its programming (Ofcom, 2017a).

Subtitle users are not a homogeneous group. They include the prelingually deaf, as well as those who lost their hearing later in life. Therefore, it is difficult to know exactly how many deaf people there are in the UK but, according to Action on Hearing Loss, over 11 million are affected by some form of hearing loss (Action on Hearing Loss, 2015).

Although subtitles are primarily created for deaf and hard of hearing audiences, they are not the only users who benefit. Hearing people may use subtitles in noisy environments, and non-native English speakers may find subtitles useful to aid comprehension (Ofcom, 2017a). In 2006, there were estimated to be over 7.5 million subtitle users in the UK (Ofcom, 2006). Recently, the ‘Subtitle it!’ campaign helped introduce new legislation, which recommends an increase in subtitled content for on-demand services (Action on Hearing Loss, 2019). This suggests that the number of subtitle users is continuing to grow.

Previous studies have found that subtitles for news programming are most frequently requested by deaf and hard of hearing viewers (Kyle, 1993; Downey, 2008). More recent findings about deaf audiences’ preferences with regards to programme content are scarce. However, broadly speaking, TV viewing figures indicate that 79% of adults still primarily get their news from television (Ofcom, 2018b; Ofcom, 2018c).

Given the range of subtitle users, opinions naturally vary on the extent to which subtitles should be edited, if at all. There are valid arguments in favour of both edited and verbatim approaches. One argument put forward is that edited subtitles prevent deaf viewers from achieving true equality of access (Robson, 2004). However, studies have shown that many prelingually deaf people have a slower reading comprehension rate compared to people exposed to spoken language (Baez Montero and Fernandez Soneira, 2010; Beal-Alvarez and Cannon, 2014). An early investigation into reading speeds for the deaf found that 145-170 words per minute was preferable (Jensema, 1997). Therefore, in order to meet these speeds, some circumstances will require subtitles to be edited to prevent latency. Delayed subtitles are cited as a source of frustration for deaf viewers (Butler, 2019).

Television news brings potential for subtitling delays and errors. Broadcasts are composed of live and pre-recorded segments (Robson, 2004). It is not possible to pre-prepare subtitles for live speech and so they must be produced in real time. Inevitably, this will cause a delay in subtitles appearing on screen. The preferred tool nowadays for subtitle production is voice recognition software, which translates the speaker's voice into text (Robson, 2004). Using voice recognition to create subtitles is faster than typing, and the software can also be trained to recognise a user's voice, thus achieving high accuracy (Dragon Speech Recognition Software, 2019). However, there are factors which can impact on accuracy, such as overlapping dialogue, unfamiliar names and trying to match the pace of fast speakers. Thus, there is a need to balance speed with accuracy (Ribas and Romero Fresco, 2008). Subtitles may be edited as a result.

Figures 1 and 2 show an extract of a weather forecast produced using voice recognition software, with mistakes and corrections highlighted in red. An error such as 'warmer' instead of 'warm air' is not especially egregious and the intended meaning can probably be inferred from context. To correct it in a live broadcast situation would cause further delay. As a result, this error may be transmitted if produced under live conditions. However, an error such as 'Maureen' instead of 'more rain' is more likely to need manual correction in order to be understood.

Subtitles differ from transcripts in that they may be paraphrased. They may also contain errors but, if transmitted, these are likely to be relatively minor. Thus, a subtitle archive would provide a reasonably accurate representation of the news content.

That is a warm front that is slowly sinking side. As its name suggests. We will go into slightly warmer tonight which means the cloud bases will be dropping down, turning misty and murky. And Maureen to the south of Iceland at the moment. That will head our way tomorrow. Again, some of that on the heavy side. We intended to fizzle out across the west this evening and tonight.

Figure 1: Extract of uncorrected weather forecast

That is a warm front that is slowly sinking south. As its name suggests. We will go into slightly warmer air tonight which means the cloud bases will be dropping down, turning misty and murky. And more rain to the south of Iceland at the moment. That will head our way tomorrow. Again, some of that on the heavy side. Rain tending to fizzle out across the west this evening and tonight.

Figure 2: Extract of weather forecast with corrections

2.6.2. Subtitles as an index

As explained in section 2.5, Snider (2000) proposed that archiving subtitles could be used as an alternative approach for viewing otherwise-inaccessible news content. However, the literature in the field of library and information science tends to focus on using subtitles as an automated index for audiovisual assets. This dates back to the 90s, when Brown *et al*. (1995) discussed synchronising subtitle files with their associated video, to facilitate retrieval system of spoken content. Heeren, Ordelman and de Jong (2008) discuss the usefulness of subtitles for automatic indexing but also acknowledge that, because subtitles are not always verbatim, they may not correspond directly with the spoken content.

Further studies use automatically generated transcripts rather than subtitles to facilitate retrieval of audiovisual material (Dowman *et al*., 2005; Ranjan, Balakrishnan and Chignell, 2006; Messina *et al.*, 2006; Demiros *et al.*, 2008). Ranjan, Balakrishnan and Chignell (2006) found some limited benefit to this, but largely concluded that the transcripts were mostly low quality, and perceived as distracting and unreadable.

Subtitles are produced using voice recognition software but, unlike the studies cited above, manual intervention ensures that the contents will be readable. I would argue that it is this manual intervention that ensures subtitles are a better representation of the audio content than automatically generated transcripts.

The studies cited in this section all focus on the use of subtitles, or automatically generated transcripts, as "collateral data" (Heeren, Ordelman and de Jong, 2008). There is little to suggest that consideration is being paid to subtitles as a valuable stand-alone resource.

2.7. Metadata

Throughout this chapter, I have focused on issues of access. In some cases, access is restricted due to barriers imposed by the institutions who own, store or maintain the material. This has led some authors, such as Kernell, Lamberson and Zaller (2017), to access the content through some form of proxy. As I argued in the introduction to this dissertation, and further discussed in section 2.6 above, I believe that subtitles could serve as a useful stand-in for the original audiovisual content. However, access to the subtitle files is also limited in that they are maintained in a "dark archive" (Allen and Johnson, 2008, p. 394) and there are no suitable mechanisms in place to help users locate content contained within documents (Riley, 2017). Thus, there is a need for rich metadata to enhance discoverability of content. This section will discuss metadata in closer detail, in order to lay the groundwork for the chapter 3, where I will explain my rationale for focusing on selected metadata elements.

2.7.1. Metadata types

The most clear and concise description of metadata types is set out in a primer by the National Information Standards Organization (NISO). This publication identifies three primary types of metadata: descriptive, administrative and structural. An additional method of is markup, in which the metadata is embedded within content itself (Riley, 2017).

Descriptive metadata contains information about the content of the document being described. Examples of elements include title, author, subject, and genre. Its primary function is as a finding aid. Administrative metadata is taken as an umbrella term for technical, preservation and rights information. It is used to help manage the resource (Riley, 2017). Structural metadata contains information about the linkages between resources. For example, a digital monograph may be composed of multiple image files, representing the front cover and individual pages. Structural metadata is applied to help understand how to navigate the resource (Miller, 2011).

Markup languages can be used alone or with other forms of metadata to denote structural features, as well as semantic information (Riley, 2017). An example of a markup language is XMLNews-Story, which uses tags in the body of news stories to identify elements such as person, location and event. It can be used in conjunction with XMLNews-Meta, which stores metadata about the news objects separately (XMLNews Technical Overview, 1999). The News Industry Text Format (NITF) follows a similar tagging strategy to XMLNews-Story, and was designed specifically to add markup to news articles (A Solution for Sharing News, 2019).

The trade-off in implementing markup in digital texts was explored in a study by Wisneski and Dressler (2009). The rationale for introducing this was as a means of improving access to the collections. However, the library had a limited number of staff and a small budget. Thus, those involved found that the markup tasks proved more time-consuming than initially anticipated.

2.7.1.1. News categorisation

Having discussed different types of metadata, I will now return briefly to the subject of vocabulary control. This was mentioned earlier, in section 2.1.3 in relation to textual news documents. However, I will now consider this in relation to broadcast news description.

Several studies applied quantitative measures when categorising news content (Amaral and Trancoso, 2003; Messina *et al.*, 2006; Pribble *et al.*, 2006). Messina *et al*. (2006) applied semantic categories with automated text classification using machine learning. No human judgement was involved, and it could therefore be argued that this is an objective way to assign subject categories. Pribble *et al*. (2006) coded health topics based on how much time was allotted to each story. These quantitative measures are in contrast with traditional cataloguing, which has a more subjective approach (Taylor, Joudrey and Miller, 2015). Subject analysis is carried out to determine the 'aboutness' of a resource. Taking a book as an example, information is deduced by looking at sections such as the title, preface, and chapter headings. Subjects are then mapped onto relevant controlled vocabulary terms (Library of Congress, 2016).

In the broadcasting industry, the IPTC maintains a list of subject terms specifically for describing news content. These terms can be used to assign broad or narrow definitions in political, cultural and sporting domains. Several studies used the IPTC controlled vocabulary, such as Allen and Schalow (1999) and Yaginuma, Mendes Pereira and Baptista (2003).

2.7.2. Metadata schemas

There is no single metadata standard that is suitable for all projects and all communities. Rather, a wide variety of metadata schemas are available, many of which are domain or media-specific (Hsieh-Yee, 2006). A metadata schema has been defined as:

> "a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of a discipline through the creation of an information system that will assist the identification, discovery and use of information within that discipline" (American Library Association Committee on Cataloging, 2000, n.p.).

Ramesh, Vivekavardhan and Barathi (2015) provide a comprehensive list of schemas in their discussion on metadata diversity. Some of these schemas are targeted to specific user communities. For example, MIDAS Heritage is used to record archaeological information, while UK LOM Core can be applied to digital objects used for educational purposes. Other schemas have more general applicability. Of these, the Dublin Core Metadata Initiative (DCMI) is known for its simplicity, flexibility and ability to be adapted to meet the needs of different communities (Haynes, 2018). For these reasons, sections 2.7.2.1 and 2.7.2.2 below will discuss DCMI alongside other schemas, which are designed for more specific resource description – namely broadcast television, and textual news content. This will provide context for methodology section 3.6, where I set out the rationale for my own metadata schema.

2.7.2.1. DCMI and PBCore

DCMI consists of 15 core elements and several additional qualifiers, all of which are optional and repeatable (Hillmann, 2005). It is a flat, rather than hierarchical, metadata model, and so does not have inherent parent-child relationships (De Sutter, Notebaert and Van de

Walle, 2006). However, the ‘Relation’ element can be used to express relationships to other resources (Miller, 2011).

One of the objectives of DCMI is to use commonly understood semantics. For example, the element ‘Creator’ could equally apply to the author of a scientific paper, or the artist responsible for the creation of a specific painting (Hillmann, 2005). However, some studies have found that semantic ambiguity and ‘fuzzy’ categories can undermine this goal (Howarth, 2003; Park and Childress, 2009).

PBCore is a metadata schema based on Dublin Core but used specifically for the public broadcasting industry (What is PBCore, n.d.). It offers element types which are clearly suited to describing television and film content, such as ‘Audience Level’ and ‘Audience Rating’. Clair (2008) found PBCore beneficial for describing the technical aspects of audiovisual collections.

Both DCMI and PBCore share the ‘Title’, ‘Subject’, ‘Description’ and ‘Contributor’ elements. However, PBCore extends this with ‘Contributor Role’. The PBCore website provides example records that highlight where this would come in useful (Sample records, n.d.). In a film with multiple cast and crew members, it is beneficial to differentiate the role that each person played in the production. DCMI does not allow for this level of description.

2.7.2.2. Application profiles

Schemas focused on textual news documents include NewsML (NewsML-G2, 2019) and NITF (A Solution for Sharing News, 2019). Yaginuma, Mendes Pereira and Baptista (2003) draw from these, as well as DCMI, to create an application profile. An application profile takes elements from different existing standards and combines them into one schema. This allows user communities to tailor existing schemas to meet their requirements (Miller, 2011).

Application profiles are used to describe resources that are not well-served by existing metadata schemas. Yaginuma, Mendes Pereira and Baptista (2003) adapted DCMI when it was unable to adequately describe their collection. A survey of 431 digital repositories found that DCMI was most commonly used as the basis for application profiles (Andrade and Baptista, 2015). However, creating an application profile can be a significant undertaking, requiring lengthy discussions in order to reach consensus about which elements to use and establish best practice guidelines (Washington and Weidner, 2017).

2.8. Ethnography

As far as the central focus of this dissertation is concerned, it is my own experience working as a subtitler that has driven this research. Chapter 3 will outline how my role has helped inform my methodology. An ethnographic approach involves the researcher observing and interacting with participants and/or data in a real-life social context. In some cases, like my own, the researcher is already a part of the context when they commence the research (Pickard, 2008).

Several studies discuss metadata creation using an ethnographic approach (Marshall, 1998; Ma, 2006; Khoo and Hall, 2013). Marshall (1998) used ethnographic data to understand the complexities involved in metadata creation for a mixed collection of physical and digital resources. Interviews revealed that staff members involved in the project had differing opinions on how to define the collection. This ambiguity had consequences for the type of metadata needed to maintain and access the resources.

A study by Ma (2006) considered the importance of ethnography in relation to the Mira Lloyd Dock collection, which involved the digitisation of glass slides featuring tree, plant and flower specimens. The user group of this collection was identified as foresters. Therefore, subject experts in the university forestry programme were consulted, who suggested using Latin plant names in the metadata descriptions as these were the terms that the user group would use to search the collection. This highlights the importance of employing the expertise of subject experts in order to create useful metadata.

Khoo and Hall (2013) identified several challenges when attempting to map the equivalent elements from their own custom metadata schema to Dublin Core. Using semi-structured interviews, they were able to link these problems to different project members' perspectives of the project. As with Marshall (1998), these differing perspectives impacted on the project and, as a result, the metadata was not applied in a consistent manner.

These studies highlight the importance of recognising how people's experiences and personal involvement can influence their decision-making with regards to metadata creation. For my own project, I will conduct the annotation alone. However, I will be doing this while paying consideration to Ob2: Create metadata for news subtitles in a way that is compatible with how it would be carried out in a professional context. That is, in a professional context, a large-scale version of this project would require the involvement of other metadata creators.

2.9. RQ1: What is the value of an archive of television news subtitles?

As technology changes and risks becoming obsolete, broadcasters are turning to digitisation to protect their assets (Wright, 2009), with the emphasis now on preserving the content rather than the carrier (Assmann and Mearns, 2015). Therefore, the value of a subtitle archive could be in providing an additional means of safeguarding the content.

It has been argued that television should be valued as an artistic creation, and not merely a document of politics, history and society (Schuller, 2015). In addition, associated written documents, such as television scripts, are collected and preserved as objects of interest (Hewett, 2014). Using this argument, subtitles are a valuable accompaniment to the broadcast footage and should be afforded respect as an integral part of the production process.

A number of sources state that TV transcripts are widely available (Snider and Janda, 1998; Robson, 2004). However, these transcripts are largely confined to the US. It is also not clear who produced them and under what conditions. Therefore, there is a need for a large-scale textual resource that comes from an unambiguous source and offers non-US content. Within the UK, subtitles are legally required for a high percentage of programming (Ofcom, 2017a). The source can also be easily traced back to the channel, or company, responsible for producing them. Manual intervention on the part of the subtitler also ensures a certain degree of accuracy. In terms of general content and widespread availability, a subtitle archive would serve as a reliable, unambiguous resource for written representations of audiovisual content.

There are several factors which could affect the value of a subtitle archive. As with the digitisation of newspapers, there is a process of remediation involved (Nicholson, 2013). The audiovisual medium is converted into a textual medium. Researchers using a subtitle archive must therefore do so with the understanding that there are differences between the two. The most obvious omission from a subtitle archive is the lack of imagery. As Maurantonio (2014) found, imagery can play a valuable role in news-based research. However, the differences between subtitles and their corresponding broadcast footage are more explicit than the differences between a print newspaper and its digital counterpart. Therefore, the limitations of a subtitle archive in relation to imagery are apparent, and the likelihood of misrepresenting the data is minimised.

Subtitles have been recognised for their value as a close approximation of speech, though not necessarily verbatim (Dowman *et al*., 2005; Heeren, Ordelman and de Jong, 2008). Additionally, subtitles may contain errors due to the conditions under which they are produced. However, this does not necessarily negate the value of a subtitle archive, provided its limitations are made explicit.

3.0. Methodology

3.1 Background to the organisation

Within the UK, there are two licence holders for the channel 3 network – ITV and STV (STV, 2018). ITV was established in 1955, broadcasting in London and soon expanding across England and Wales. Since 2016, it has also maintained control of programming in Northern Ireland (About ITV, 2019). STV began as Scottish Television in 1957, and now reaches over 3.5 million viewers a month (STV, 2017). It operates throughout most of Scotland, apart from the border region between Scotland and England, which is managed by ITV (Ofcom, 2017b). There is a professional relationship between ITV and STV, with shared content and advertising (McIvor, 2012). However, organisationally and legally speaking, they are independent companies.

STV maintains its own distinct localised news coverage to four separate broadcasting regions – north, east, west and Tayside. There is no south region because, as discussed above, this is controlled by ITV. These regions are more commonly referred to by the cities in which the studios are based – Aberdeen (north), Edinburgh (east), Glasgow (west) and Dundee (Tayside). To reflect typical usage in my working life, I will henceforth use the city names to refer to the four regions in which STV operates.

During the regular working week, there are lunchtime, early evening and late evening news bulletins. Until recently, STV had a strong focus on news and delivered well in excess of its required output (Ofcom, 2018a). From 2011 until 2018, there were three full length 6pm programmes for Aberdeen, Edinburgh and Glasgow. Within the Aberdeen programme, there was also a short opt-out for Dundee. However, in September 2018, a restructuring process within the company brought a change in format. Aberdeen and Dundee were unaffected, but the Glasgow and Edinburgh programmes merged to create a single 6pm news bulletin, broadcasting throughout Scotland's central belt. This new programme contains short opt-outs of more localised content, transmitted separately but simultaneously to Glasgow and Edinburgh (STV, 2011; STV, 2018).

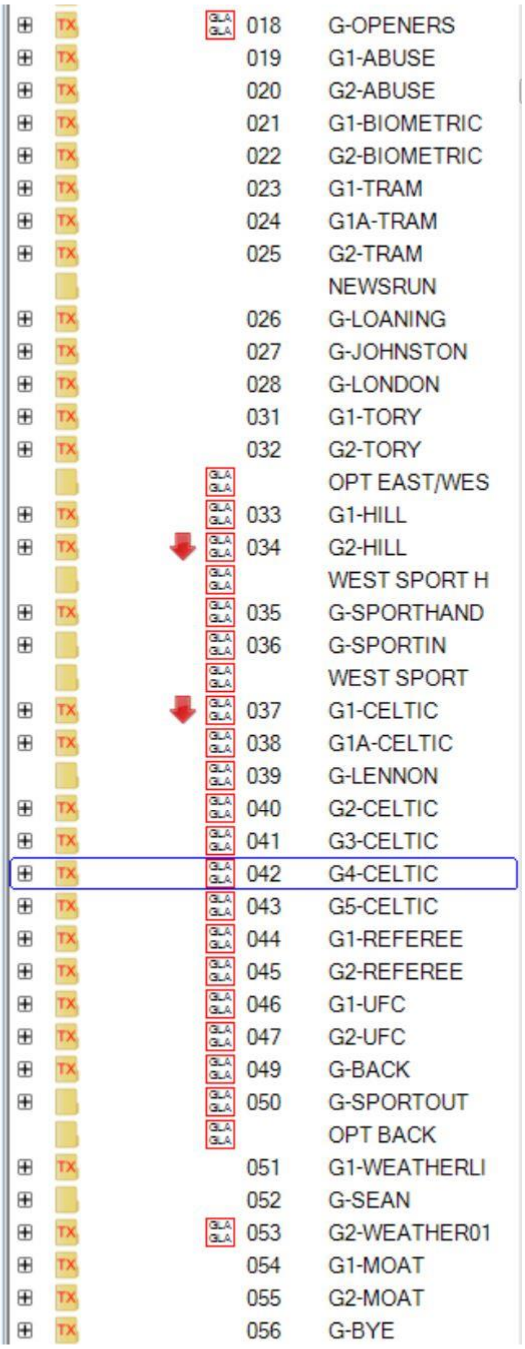
3.2. News terminology

As its name suggests, a news running order is an account of all the content of the programme, displayed in the sequence in which it will be transmitted. Figure 4 shows an example of an STV news running order. As this contains technical instructions and nontranscribed segments that are irrelevant to this discussion, a more user-friendly version of the running order is shown in figure 5.

Contained within the running order are the individual items which make up the content. The term 'item' is used to refer to a discrete digital object (Miller, 2011). Figure 3 below explains the terminology used to refer to these items. These definitions are taken from Glossary of Broadcast Terms (2011), as well as my own understanding based on experience.

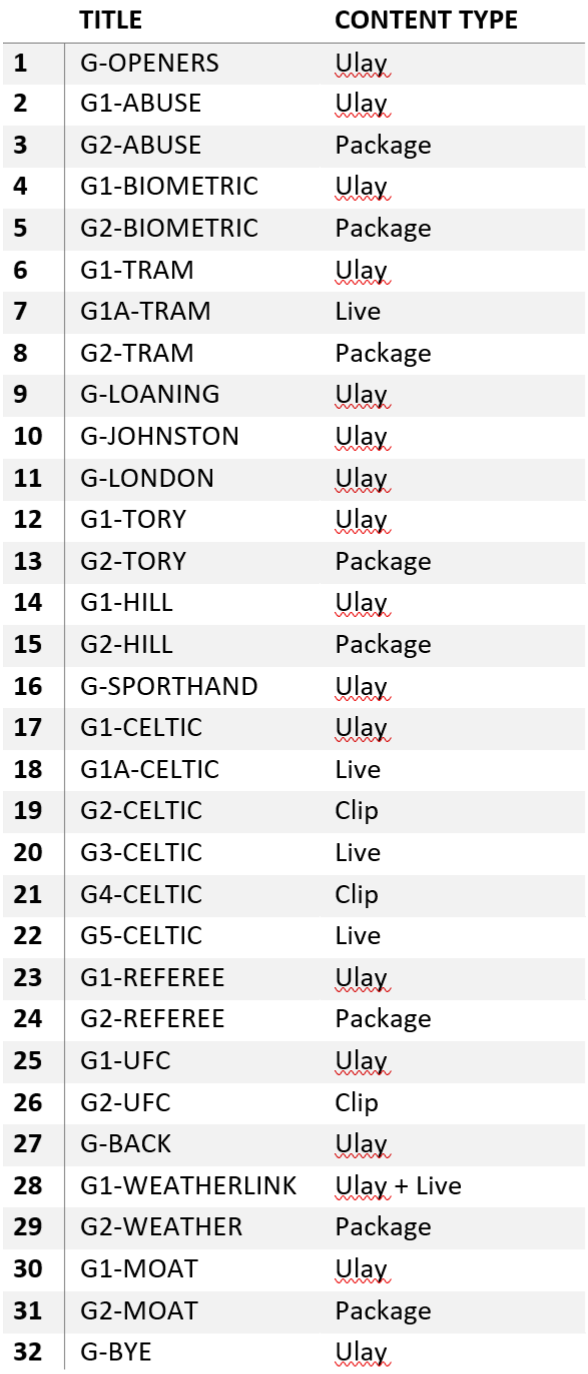
ITEM NAME	DESCRIPTION
UNDERLAY (ULAY)	Words read by the news anchor from an autocue script; can also contain unscripted ad-libs.
CLIP	A soundbite featuring one or more speakers; short in duration (typically under one minute).
PACKAGE	A longer report containing edited material; combines clips, interviews, voiceover and pieces to camera.
LIVE	Broadcast in real-time, often on location; reporters may pre-prepare a script from which to read but typically the item is unscripted.

Figure 3: News terminology and definitions



⊞	TX	018	G-OPENERS
⊞	TX	019	G1-ABUSE
⊞	TX	020	G2-ABUSE
⊞	TX	021	G1-BIOMETRIC
⊞	TX	022	G2-BIOMETRIC
⊞	TX	023	G1-TRAM
⊞	TX	024	G1A-TRAM
⊞	TX	025	G2-TRAM
			NEWSRUN
⊞	TX	026	G-LOANING
⊞	TX	027	G-JOHNSTON
⊞	TX	028	G-LONDON
⊞	TX	031	G1-TORY
⊞	TX	032	G2-TORY
			OPT EAST/WES
⊞	TX	033	G1-HILL
⊞	TX	034	G2-HILL
			WEST SPORT H
⊞	TX	035	G-SPORTHAND
⊞		036	G-SPORTIN
			WEST SPORT
⊞	TX	037	G1-CELTIC
⊞	TX	038	G1A-CELTIC
		039	G-LENNON
⊞	TX	040	G2-CELTIC
⊞	TX	041	G3-CELTIC
⊞	TX	042	G4-CELTIC
⊞	TX	043	G5-CELTIC
⊞	TX	044	G1-REFEREE
⊞	TX	045	G2-REFEREE
⊞	TX	046	G1-UFC
⊞	TX	047	G2-UFC
⊞	TX	049	G-BACK
⊞		050	G-SPORTOUT
			OPT BACK
⊞	TX	051	G1-WEATHERLI
⊞		052	G-SEAN
⊞	TX	053	G2-WEATHER01
⊞	TX	054	G1-MOAT
⊞	TX	055	G2-MOAT
⊞	TX	056	G-BYE

Figure 4: Typical STV news running order



	TITLE	CONTENT TYPE
1	G-OPENERS	Ulay
2	G1-ABUSE	Ulay
3	G2-ABUSE	Package
4	G1-BIOMETRIC	Ulay
5	G2-BIOMETRIC	Package
6	G1-TRAM	Ulay
7	G1A-TRAM	Live
8	G2-TRAM	Package
9	G-LOANING	Ulay
10	G-JOHNSTON	Ulay
11	G-LONDON	Ulay
12	G1-TORY	Ulay
13	G2-TORY	Package
14	G1-HILL	Ulay
15	G2-HILL	Package
16	G-SPORTHAND	Ulay
17	G1-CELTIC	Ulay
18	G1A-CELTIC	Live
19	G2-CELTIC	Clip
20	G3-CELTIC	Live
21	G4-CELTIC	Clip
22	G5-CELTIC	Live
23	G1-REFEREE	Ulay
24	G2-REFEREE	Package
25	G1-UFC	Ulay
26	G2-UFC	Clip
27	G-BACK	Ulay
28	G1-WEATHERLINK	Ulay + Live
29	G2-WEATHER	Package
30	G1-MOAT	Ulay
31	G2-MOAT	Package
32	G-BYE	Ulay

Figure 5: User-friendly running order

3.3 Subtitling at STV

Along with the four other terrestrial channels in the UK, STV is responsible for public broadcasting in relation to its news content. Therefore, in accordance with Ofcom statutory requirements, 90% of its televised content must be subtitled (Ofcom, 2017a; Ofcom, 2018a). Along with my colleagues, I produce the subtitles for all STV news bulletins.

For most items in the running order, subtitles can be pre-prepared. For example, the autocue script, from which the anchor reads, is generated approximately 90 minutes before on-air time. However, due to the fast-paced nature of the newsroom, the content is frequently updated both in the run-up to and during transmission. As shown in figure 4, red arrows indicate that there has been an update. Using specialist subtitling software, I have direct access to the newsroom running orders, which automatically update to reflect any changes that are made (Screen Systems, 2015).

Subtitles also need to be produced for clips and packages. These are more time consuming to prepare as no autocue script is available. Thus, they need to be created completely from scratch. On occasion, reporters include a rough voiceover script for their packages, but these need to be checked against the audio for accuracy and completeness. STV uses the MediaCentral asset management platform to manage its news content (MediaCentral, 2019). From here, I can access clips and packages when they have been uploaded by the reporter responsible for creating and compiling them. Although this should be done in a timely manner, the reality of working in a fast-moving industry with tight turnaround times means that the content is sometimes not uploaded until we are already on-air. Consequently, any packages or clips that are received late mean that subtitles must be produced live, which can impact on the quality. For more discussion on quality issues, see section 2.6.1.

3.4. Design

My approach to this research follows a case study design of STV news subtitles, exploring how the use of metadata can help enhance discoverability of content. This study is exploratory and interpretive in nature. I am employed by STV as a subtitler, and am one of a small team of people responsible for producing the output which I intend to analyse. Using my personal knowledge and experience, I will also draw from the field of ethnography to help inform my research.

There are similarities between ethnography and case study. However, ethnography is concerned with describing and interpreting a cultural phenomenon, while case study is focused on developing in-depth information about a phenomenon through analysis of individual cases (Pickard, 2008). It is possible to apply an ethnographic approach within a case study. Several authors using this approach have shown that insider knowledge can play a valuable role in interpreting how resources are being used and thus how they can be best managed (Atton, 1998; Ma, 2006). I have no formal media training, but am embedded within a news organisation. I help to create news output in the form of subtitles. Due to my role, I have developed an understanding of broadcast news operations specifically at STV. I can use this insider knowledge to inform my choice of subtitle metadata.

My primary research method is a case study because I am seeking to describe the possibility of a subtitle archive within one organisation. Other broadcasting companies employ the services of subtitlers and have different procedures, responsibilities and demands. For these companies, the realities of such an archive will be different. Data from a case study is not typically used to make generalisations as it focuses on a single phenomenon (Pickard, 2008). I cannot therefore comment on the applicability of a subtitle archive outwith the context of STV.

3.5. Data collection and sampling

The STV news subtitle archive dates back to 2009. Due to my employee status, I was able to access this primary source data. The files can only be opened using specialist subtitling software and, as such, had to be viewed on site. My sample consists of two news running orders. In order to achieve Ob3, as set out in section 1.2, I used a historic dataset from 2014 and a more recent dataset from 2019. To gather the data, I used a purposeful sampling technique. This is typically used in qualitative research to select information-rich sources which help gain a comprehensive understanding of issues relevant to the inquiry (Palinkas *et al.*, 2015).

Ob2 of this research is to explore how the creation of metadata can provide added value to subtitle files. Subtitles are intended to be viewed alongside visual content. If seen in isolation, it is therefore inevitable that some contextual information will be missing from subtitles that would otherwise have been provided by the video, such as speaker identification. Inserting this missing contextual information through the metadata records will help provide added value. In order to achieve this, I must view the associated video files. However, these videos are only available through MediaCentral for approximately two months after their original air date. As a result, any content older than this is inaccessible to me. Thus, a historic subtitle sample can only be viewed in isolation. By comparing the 2014 and 2019 running orders, I can identify what challenges arise because of the different level of access to video content, thus enabling me to answer RQ3, as outlined in section 1.1.

The news subtitle archive is organised by time, region and date. There is no way of knowing the content of each running order without opening the files using specialist subtitling software and reading the individual news stories. This is time-consuming and limited my ability to choose a purposeful sample. I selected samples according to the time, region and date that I felt would be most likely to offer information-rich content. This follows the critical case strategy, in which a small and carefully selected sample can be used to "yield the most information and have the greatest impact on the development of knowledge" (Patton, 2002, p. 236). My rationale for the criteria is set out below.

3.5.1. Time

I used running orders from the 6pm evening news because this slot has the highest viewer share (Ofcom, 2018a). At approximately 28 minutes, it is also longer than the lunchtime and late-night news bulletins. This means I will have a greater number of items to analyse than a sample taken from one of the other bulletins. In addition, research has found that this is the time when viewers are most engaged and focused on the news (Ofcom, 2018d). This suggests that the 6pm evening news will provide the most information-rich content.

3.5.2. Region

Ideally, both samples would be taken from the Glasgow news bulletin. The STV headquarters are in Glasgow and this is where most of the reporters are based. Therefore, there is a sense that the Glasgow news is the flagship programme. In my personal experience, the type of stories that are broadcast in Glasgow are varied in terms of content. In contrast, the Aberdeen programme typically devotes a large proportion of the news to stories about the oil and gas or fishing industries. These are more tailored to localised interests of audiences in the north of Scotland. In contrast, the Glasgow news programme seems to have more general applicability.

However, due to the 2018 STV news restructuring process, there is no longer a Glasgow bulletin. The 6pm evening news is now broadcast simultaneously to Glasgow and Edinburgh and can be more accurately termed the central news. In practice, this functions in the same way as the Glasgow news bulletin which ran from 2011-2018. It is still the flagship programme and features the same kind of content that it did prior to its merger with Edinburgh. The main difference is that it is now co-anchored rather than presented by a single person. As explained in section 3.1, there is also a short opt-out for both cities. Essentially, this means that for approximately five minutes, viewers in Glasgow will be seeing different content from viewers in Edinburgh. My 2019 sample will be taken from the central news bulletin and will exclude the Edinburgh opt. My 2014 sample will be taken from the Glasgow news bulletin.

3.5.3. Date

My 2019 sample was limited to the months of May and June, as these were when the associated videos were available for me to view. As I was using an ethnographic approach, I had to be cautious not to intervene during the data collection process in a way that might bias the results (Khoo and Hall, 2013). For that reason, I took a sample of news that I was not responsible for creating or transmitting. Rather, one of my colleagues created and transmitted the subtitles for that bulletin.

For my historic sample, I chose a date prior to September 2014. This seemed like it would provide information-rich content as this was in the run-up to the Scottish independence referendum. Therefore, it was probable that there would be several news stories devoted to this emotive topic. This is also a topic of interest in 2019 due to Brexit. Therefore, I felt the samples from 2014 and 2019 would be equally balanced in terms of political representation.

3.6. Metadata schema

News subtitles have characteristics in common with both digitised newspapers and broadcast news. As a result, I could not find an existing metadata schema that met all my requirements. However, schemas like XMLNews and NITF, which are designed for news texts, contain many elements that would be irrelevant to my datasets. On the other hand, a schema like PBCore, which is designed for public broadcasting, contained some useful elements but did not differ significantly from DCMI (What is PBCore, n.d.).

As discussed in section 2.9.4, Yaginuma, Mendes Pereira and Baptista (2003) created an application profile. This was based on DCMI but also contained elements taken from other schemas. As I could not find a metadata schema that met my requirements, I followed a similar example and created my own application profile. I used DCMI and the Dublin Core qualifiers (DCQ) as the primary schema, but drew from other sources when necessary.

For this study, I am interested in resource discovery and have therefore mostly focused on descriptive metadata. However, I also include some structural metadata elements. The physical layout of newspapers has been found to be important for research purposes and structural metadata has been applied to account for that (Allen and Schalow, 1999). It may therefore be of interest to external users to understand how the various news items are linked together in broadcast news.

Issues around copyright and intellectual property have been raised with regards to a news subtitle archive (Snider, 2000). However, a full discussion of this lies beyond the scope of this dissertation. Therefore, I will disregard rights metadata in my schema.

Figure 6 shows my element set, including the source they are derived from, encoding schemes, and further notes to clarify usage. My reasons for choosing some elements require clarification. This is outlined below.

Edition: This element was taken from NITF, which is used to describe news texts. However, for my datasets, it is used to show the edition of the news programme from which the subtitles are taken.

Region: As above, NITF provides elements for publication information, such as the regions in which newspapers are published. For my datasets, it will refer to the regions in which the subtitles were broadcast.

Subject: For the subject element, I allowed for the possibility of using both controlled vocabulary and free-text keywords. In order not to confuse them, these are entered into separate subject iterations. The first iteration is reserved for controlled vocabulary terms, taken from the IPTC Media Topics taxonomy. This is an updated version of the Subject Codes taxonomy, as used in several studies cited throughout section 2 of this dissertation, such as Allen and Schalow (1999). As discussed in section 2.7.1.1, Media Topics is designed for describing news media. As my datasets contain news media content, Media Topics seemed like a suitable choice for controlled vocabulary. The second iteration is reserved for free-text keywords.

Creator/Creator Role: DCMI has a creator element, but it does not have a way of indicating the role of this person in the creation of the resource. There are multiple people involved in the production of broadcast news. Therefore, I felt it would be informative to list not only their names, but also give some indication of their involvement. PBCore provided this option, as well as an associated controlled vocabulary to describe roles within the television industry. They did not provide a term for news anchor. However, 'commentator' seems to fit the definition of what the news anchor does and so this was used instead (PBCore vocabularies, n.d.). I considered the use of LCNAF to encode personal names, but a preliminary search of names from my datasets did not turn up any LCNAF entries. I took this as an indication that many of the named people in my dataset would not be well known enough to have LCNAF entries. Thus, personal names are entered using free-text keywords. In sample records on the PBCore website, personal names were listed forename first, followed by family name. This is in contrast to guidance given by DCMI, which states that the order should be inverted (Hillmann, 2005). I felt it would be more intuitive to list names in the order in which they are normally spoken and thus, I followed PBCore's example with regards to this.

Contributor/Contributor Role: I took these from PBCore for the same reason explained above. This is intended to be used to describe people who provide soundbites to clips or packages.

Has Part/Is Part Of: Some items in the running order are related, in that they are about the same story and should be viewed in a particular order. Thus, these elements are used to indicate how one item naturally follows on from another. For example, as shown in figure 5, some items have the same title but a different initial number. For example, G2- BIOMETRIC follows on from G1-BIOMETRIC.

Name	**Definition**	**Source**	**Encoding scheme**	**Notes**
Title	Name given to the resource	DCMI		Take from title in running order, e.g. G4-CELTIC
Identifier	Unambiguous reference to the resource	DCMI		Edition, region, date, title of item, e.g. NEWSATSIXCEN310519G4CELT
Date	Date of transmission	DCMI	ISO 8601 YYYY-MM-DD	
Edition	Edition of programme	NITF		STV News at Six
Region	Geographic area	NITF		Where the programme is being transmitted
Description	Free text account	DCMI		
Subject	Topic(s) of the resource	DCMI	IPTC NewsCodes Media Topics	Separate multiple entries with a semi-colon
Creator	Primary person responsible for creating resource	PBCore		Name of reporter or presenter
Creator Role	Role of the creator	PBCore	PBCore Vocabulary	Reporter, Speaker, Interviewer, Interviewee, Commentator
Contributor	Person responsible for making contributions	PBCore		Names of notable people, e.g. politicians, sportspeople who feature in the clip or package
Contributor Role	Role of the contributor	PBCore	PBCore Vocabulary	Reporter, Speaker, Interviewer, Interviewee, Commentator
HasPart	Resource logically includes referenced resource	DCQ		Refer to title of item
IsPartOf	Resource is a logical part of referenced resource	DCQ		Refer to title of item

Figure 6: Metadata schema

3.7. Procedure

As discussed in section 2.3.2, I followed the all-in approach to preservation (Balbi, 2011). Therefore, for both samples, I created a metadata record for every item in the news running order. The 2019 dataset contained 32 items (as shown in figure 5), while the 2014 dataset contained 40 items. The intended users of a subtitle archive are difficult to identify at this preliminary stage. Therefore, I cannot judge what may be valuable or useful from the perspective of external users, providing further justification to an all-encompassing approach to metadata creation.

Figure 7 shows a blank metadata record, which I used to enter the values for my annotation. As Miller (2011) observes, metadata creators typically enter values in a user-friendly interface, rather than directly code the metadata in XML. Figure 7 was intended to resemble a user-friendly interface, in order that my process be in line with Ob2 in section 1.2. That is, I wanted to enter the metadata in a similar manner to how the process would be conducted in a professional setting.

TITLE	
IDENTIFIER	
DATE	
EDITION	
REGION	
DESCRIPTION	
SUBJECT	
SUBJECT	
CREATOR	
CREATOR ROLE	
CONTRIBUTOR	
CONTRIBUTOR ROLE	
HASPART	
ISPARTOF	

Figure 7: Blank metadata record

As discussed in section 2.7.1.1, some studies applied quantitative measures to categorise news data (Pribble *et al.*, 2006; Messina *et al*., 2006). In contrast, I used a qualitative approach for my analysis. This also relates to Ob2 of my research. In my everyday working life, I do not have access to the kind of tools needed to conduct quantitative analyses on news data. Therefore, quantitative measures would not be compatible with how the creation of subtitle metadata would be carried out in a professional context.

I followed a traditional cataloguing approach to determine the aboutness of the resource (Library of Congress, 2016). I then mapped the concepts to controlled vocabulary terms from the IPTC Media Topics vocabulary. On occasions where appropriate controlled vocabulary could not be assigned, as discussed in section 3.6, free-text keywords were entered into a separate subject iteration.

In total, the 2019 dataset took three hours to annotate, while the 2014 sample took two hours and 15 minutes. I had access to the corresponding video footage for the 2019 dataset, but not the 2014 dataset. The different levels of access to video footage enables me to carry out Ob3 of my research, as outlined in section 1.2. Comparisons between recent and historic data will be drawn from the fact that historic subtitle files can only be viewed in isolation.

4.0 Analysis

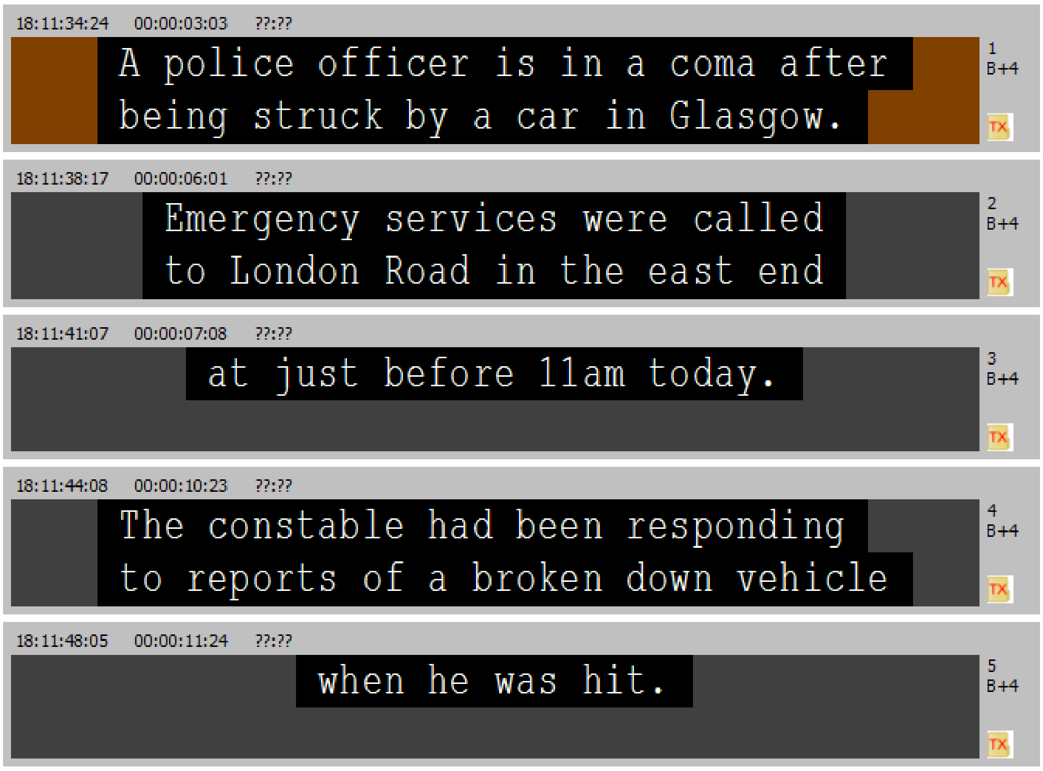
In this chapter, I will address the research questions outlined in the introduction to this dissertation, section 1.1. As also discussed in section 1.1, this research is intended as an exploratory account of the benefits and challenges of an accessible subtitle archive. Chapter 2 discussed the benefits. This chapter will therefore mainly focus on problem areas. I identify four main problem areas – aboutness and translation, free-text keywords, item-level metadata, and speaker identification. However, for the sake of clarity, these main problem areas have also been further subdivided where necessary.

Due to limitations of space, it is not within the scope of this dissertation to analyse every problem area to the same degree of depth and detail. However, it is hoped that even a brief discussion will offer useful insights into the process of establishing and maintaining an accessible subtitle archive.

4.1. Aboutness and translation

As set out in methodology section 3.7, I employed the traditional bibliographic method of subject analysis to determine the content of each item (Library of Congress, 2016). In terms of subtitle files, there are few clues to assist with determining the aboutness of the items. Beyond the words contained within the subtitle files themselves, the only additional clue offered is a title. However, these titles are assigned by the editorial news staff for the purpose of differentiation. Therefore, a title cannot be taken as a reliable indicator of content. In other words, the titles are either too generic to provide any useful context, or in other instances, can be misleading.

For example, figure 8 shows the subtitles for an item with the title ‘London’. However, as the content shows, the news item concerns a street in Glasgow called London Road, rather than the city of London. Thus, as there were no reliable indicators of content available to use as clues, I had to skim-read the content to ensure I did not misrepresent the aboutness of each item.



18:11:34:24 00:00:03:03 ??:??

A police officer is in a coma after
being struck by a car in Glasgow.

18:11:38:17 00:00:06:01 ??:??

Emergency services were called
to London Road in the east end

18:11:41:07 00:00:07:08 ??:??

at just before 11am today.

18:11:44:08 00:00:10:23 ??:??

The constable had been responding
to reports of a broken down vehicle

18:11:48:05 00:00:11:24 ??:??

when he was hit.

Figure 8: Subtitles for G-LONDON

After aboutness had been determined, I selected terms from the Media Topics controlled vocabulary to enter as values into the 'subject' element of my metadata schema. For a reminder of why I chose the Media Topics controlled vocabulary, see methodology section 3.6. For this stage, Joudrey, Taylor and Miller (2015) suggest creating an aboutness statement to describe the subject matter, consisting of a few sentences or a paragraph. I did not do this as it would seem to undermine Ob2: Create metadata for news subtitles in a way that is compatible with how it would be carried out in a professional context.

Creating an aboutness statement for every individual item would not be practical in a professional context, due to the large number of items that need to be analysed. I did, however, note down some words and themes which I used as a starting point for the translation phase.

4.1.1. Analysis of aboutness and translation

4.1.1.1. Aboutness

My expectation was that my level of familiarity with the stories contained in the datasets would have a significant impact on my ability to determine the aboutness of each item. That is, I expected the recent 2019 dataset would be straightforward to analyse, while the historic 2014 dataset would present more challenges. This is because coverage of news stories can span several days, or longer. Therefore, even though I did not originally create the subtitles used in the 2019 sample, my position at STV means that I have a general sense of awareness of news stories in any given week. In contrast, the 2014 dataset was taken from a time prior to my employment at STV. I did not expect the stories to be familiar. In addition, I also was unable to access the historic video footage. Thus, I had no supplementary contextual information to consult and had only the subtitle files themselves to interpret.

However, in practice, there was little difference in the process of determining aboutness between recent and historic datasets. This can be attributed to the fact that television news broadcasts are limited by their format. The stories contained within a bulletin are designed to hold viewers' attention and therefore, each segment tends to be a short summary of events rather than an in-depth investigation (Murphy, 1997). For a reminder of news terminology, see figure 3 in section 3.2. Ulays, as exemplified in figure 8, are typically very short. Containing only five subtitles, the content is not long enough to be overly complex. As a result, even when the stories were unfamiliar to me, they did not take long to read and analyse.

Package items are longer than Ulays (for an example of a package, see appendix B). As a result, the content takes longer to read and does provide more detail than a Ulay. However, each package is preceded by an accompanying Ulay, which functions as a short introduction to the package. Thus, the accompanying Ulay item allowed me to gain a sense of what the package contents. In most instances, I identified the same themes in Ulays and packages. As shown in figures 9 and 10, this results in considerable duplication of subject terms in individual metadata records. This matter of duplication will be addressed in sections 4.1.2.1 and 4.4.

8

TITLE	G1-RALLY
IDENTIFIER	NEWSATSIXGLA020614G1-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Scotland’s top law officer has refused to rule out a criminal prosecution over death of spectators at a car rally in the Borders
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Scottish Borders
CREATOR	John Mackay
CREATOR ROLE	Commentator
HASPART	G2-RALLY
	G3-RALLY
	G4-RALLY
	G5-RALLY
	G-WEBSITE

Figure 9: Metadata record for Ulay

9

TITLE	G2-RALLY
IDENTIFIER	NEWSATSIXGLA020614G2-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Close friends and neighbours speak about those who died at the car rally tragedy in the Scottish Borders
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Scottish Borders
CREATOR	Sharon Frew
CREATOR ROLE	Reporter
ISPARTOF	G1-RALLY

Figure 10: Metadata record for package

As mentioned in methodology section 3.7, the annotation process took slightly longer for the 2019 dataset, which I attribute to the time taken to re-watch video footage. In contrast, I was unable to re-watch video footage for the historic 2014 dataset and thus, the process took less time overall. In section 2.1.4, I discussed the role of visual imagery and how the absence of this can lead to text being misinterpreted. The lack of visual clues did present challenges with speaker identification for the 2014 dataset, but I would argue that it did not impact on determining aboutness. However, I will return to the problem of speaker identification in section 4.4.

Although there is video footage available for recent data, I do not believe this provides much additional benefit in terms of determining aboutness. If anything, relying solely on the subtitle files is more efficient, which relates back to RQ4. In a real-life context, re-watching video footage results in a greater trade-off in terms of time, with little extra reward.

With regards to RQ3, the first phase of the subject analysis process – determining aboutness – can be carried out on both historic and recent data without much difficulty. Although there is less familiarity with the content of historic data, the short, summarised nature of broadcast news ensures that the overall themes and concepts can be easily extracted. However, regarding the second phase of the subject analysis process – translation – I encountered several challenges, which I will discuss in section 4.1.1.2 below.

4.1.1.2. Translation

As discussed in literature review section 2.1.3, controlled vocabularies have strengths and weaknesses. Thus, I anticipated that Media Topics might not meet all my requirements, which I was why I also planned to use free-text keywords in a separate iteration of the subject element. However, Media Topics is specifically designed for the purpose of describing news content, and therefore my initial expectation was that, overall, it would be suitable.

However, I found applying the subject terms to be challenging. As discussed in section 4.1.1.1, the content is often very short. While this was advantageous in terms of determining aboutness, it did not provide much content to analyse and therefore it could be challenging to assign more than one subject term. On several occasions, I only applied one subject term to my metadata records from the Media Topics taxonomy.

Another problem was that Media Topics seemed to be lacking in specificity for certain domains, a problem highlighted by Macgregor and McCulloch (2006). For example, figure 11 shows the metadata record for an item about war. There was no term in Media Topics to indicate what war was being referred to. The vagueness of the term ‘war’ felt like it was no useful for resource discovery. I did, however, supplement this with free-text keywords, as seen in the second ‘subject’ element iteration. I will discuss free-text keywords in more detail in section 4.2.

38	
TITLE	G1-BROTHERS
IDENTIFIER	NEWSATSIXGLA020614G1-BROTHERS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Descendants of brothers who fought in WWI commemorate their sacrifices
SUBJECT	War; ceremony
SUBJECT	WWI; World War 1; First World War; Great War
CREATOR	John Mackay
CREATOR ROLE	Commentator
HASPART	G2-BROTHERS

Figure 11: Metadata record for G1-BROTHERS

Another issue I encountered was that Media Topics seems more targeted towards US resources. Although I cannot find anything that explicitly says this is the case, it is well documented that controlled vocabularies are not neutral, and reflect cultural biases (Olson, 2001; Strottman, 2007; Ragaller and Rafferty, 2012; Library of Congress, 2016). However, there is little in the scholarly literature that discusses this with regards to Media Topics or its predecessor, Subject Codes. Although a detailed examination of this controlled values and their inherent biases is outside the scope of this dissertation, there were certain domains in which this stood out. For example, terms such as ‘campaign finance’, ‘primary elections’ and ‘citizens initiative and recall’ do not apply to UK politics, suggesting that non-US resources might not be well served by Media Topics.

As discussed in methodology section 3.5.3, I anticipated the datasets to feature stories about Scottish independence and Brexit, thus providing me with information-rich content to analyse. Yet without the appropriate terms to describe this content, it felt like the items were not being sufficiently described in a way that would be useful for people searching for this kind of content. Media Topics has the term ‘referenda’, but does not provide anything more specific than that.

Another notable example where Media Topics did not meet the requirements of my data was found in the sports domain. Sports in general are very well represented in the Media Topics taxonomy, with 104 terms for individual competition disciplines. However, the term ‘soccer’ is used instead of ‘football’, which felt jarring to use in a UK context

This raises the issue of warrant. There are three main types of warrant which guide the creation of controlled vocabularies: literary, user and organisational. Literary warrant justifies the inclusion of terms based on how they are used normally in the resources. User warrant favours terms based on what users prefer. Organisational warrant refers to the terms used by the organisation that will use the controlled vocabulary (National Information Standards Organization, 2010). Of these three types, I will disregard organisational warrant from further discussion. The needs of the organisation are not relevant to this dissertation, because the intended users of a proposed subtitle archive would be external to the company.

However, there is tension between literary warrant and user warrant. The topic of football occurs frequently in STV News. I know this from my own experience working within the organisation. It is also highlighted in the data samples. In the 2014 dataset, 7 individual items have some connection to football. In 2019, this number rises even higher to 11 items. Therefore, within the data, ‘football’ is the preferred term.

Regarding user warrant, as discussed in 2.5.2, a limitation of existing transcript archives is that they are largely confined to the US. This highlights a need for a non-US source of news content for people to access. However, just because the source of news is Scottish, as my proposed subtitle archive would be, does not imply that the users accessing it are based in

Scotland or even the UK. As stated in methodology section 3.7, it is anticipated that the user group is far-reaching. The term ‘soccer’ may therefore appeal to a larger user base.

The current status of the subtitle archive is not well served for discoverability. As discussed in section 3.5, users are limited to date, time and region. Beyond these criteria, finding content is largely down to luck and guesswork. Therefore, with regards to RQ2, even the application of a small number of subject terms would enhance discoverability. Media Topics does not always allow for a rich level of indexing, but still offers an improvement on what is currently available.

4.1.2. Suggestions to improve aboutness and translation

While determining aboutness was straightforward, the translation phase presented some difficulties because of a lack of suitable subject terms in the Media Topics controlled vocabulary. This may be due to limitations of the controlled vocabulary itself (Joudrey, Taylor and Miller, 2015). However, all controlled vocabularies have drawbacks. For a fuller discussion of controlled vocabularies, see section 2.1.3. Therefore, the sections below offer some suggestions for improvement. Section 4.1.2.1 is a suggestion that works within the limitations of Media Topics, while section 4.1.2.2 considers whether alternative encoding schemes would be more suitable for my datasets.

4.1.2.1. Intercoder reliability testing

Krippendorff (1980) first introduced the concept of intercoder reliability, in which independent judges apply the same coding system to data, in order to determine if they reach the same decisions. This helps to ensure a level of consistency is applied across metadata records. Other studies previously cited in the literature review, such as Pribble *et al*. (2006) and Hale, Fowler and Goldstein (2007), conducted intercoder reliability testing and found a high level of agreement between coders for their respective assigned categories.

In contrast, Maguire (2002) did not use intercoder reliability testing. This might lead to greater consistency initially, as only one person has the responsibility of creating the metadata. However, the disadvantages are that it may result in bias and oversight. As discussed in section 3.7, I relied on my own judgement for this process and therefore I was the sole coder. Within a professional context, intercoder reliability could achieve more credible results. In order to ascertain how this might be done, I conducted a small-scale experiment wherein I asked a subtitling colleague, Alex[1], to apply Media Topics terms to six news items that I pre-selected. I selected a mixture of items from both the 2014 and 2019 datasets. Some were selected because I felt they were straightforward in terms of analysis, while others were longer and more complex. I also included some items that consisted of a Ulay and its associated package, and asked him to consider them as separate items, as I had done myself.

In order to ensure Alex approached the data in a way that was commensurate with my own methodological approach, the instructions I provided him with were quite broad (see appendix A). All I asked was that he assign more than one category to each item and that he not spend an excessive amount of time carrying out the task. Figure 12 below shows his results alongside my own. The words highlighted in red show where we both applied the same subject term for the same item.

[1] This is a fictional name assigned to ensure anonymity

	TITLE	MY SUBJECT TERMS	ALEX'S SUBJECT TERMS
SUBJECT	G-JOHNSTON	Homicide; monument and heritage site	Crime, law and justice; crime; homicide; law enforcement; investigation (criminal)
SUBJECT	G1-TENNIS	Tennis; international games	Sport; competition discipline; tennis; sport event; international cup
SUBJECT	G1-POLL	Referenda; political campaigns; political parties and movements	Politics; election; referenda; political process; political parties and movements
SUBJECT	G2-POLL	Referenda; political campaigns; political parties and movements; demographics; gender	Politics; election; referenda; political process; political parties and movements; society; mankind; gender
SUBJECT	G1-BIOMETRIC	Police; personal data collection; identification technology	Crime; law and justice; law enforcement; police; science and technology; technology and engineering; identification technology
SUBJECT	G2-BIOMETRIC	Police; personal data collection; identification technology	Crime; law and justice; law enforcement; police; science and technology; technology and engineering; identification technology

Figure 12: Intercoder reliability testing results

It is important not to generalise too much from these results, as I only asked one colleague to participate, and the data sample was small. However, the subject terms that Alex selected do seem to suggest considerable agreement with my own chosen terms. He applied a greater number of terms than I did, which can be attributed to several factors. Firstly, he was only required to choose subject terms for 6 items, as opposed to my 72 items. Therefore, I was perhaps less inclined to spend the same amount of time as Alex did when choosing terms for these items.

Another reason is that he included both broader and narrower terms from the taxonomy, while I only included the narrower terms. This relates to the principle of specific entry, in which a resource should be described using the most precise term offered by the controlled vocabulary (Joudrey, Taylor and Miller, 2015). My instructions were admittedly left quite open, perhaps causing him to feel uncertain about this. Additionally, there is nothing stated on the Media Topics taxonomy which states you should only apply the most specific term. In a real-life setting, this information could be clarified in a policy statement and best practice guidelines.

Even when subject terms did not match exactly, there was still some similarity in meaning in the terms we selected. For example, G2-POLL is about women voting. In my case, I chose the term 'demographics' while Alex chose 'society' and 'mankind'. However, we both identified that the story concerned characteristics of groups of people, and matters pertaining to the population.

This may be down to subjective nature of subjective nature of subject analysis. However, Suominen and Mader (2013) conducted a quality assessment on 24 controlled vocabularies, including Media Topics. Although they did not provide specific examples, they indicated that Media Topics had issues with undocumented concepts and overlapping labels. Thus, the fact that we chose similar terms may be because there were multiple viable options available.

Intercoder reliability testing suggests that Media Topics can be used successfully for my datasets, with agreement between independent coders showing consistency in application. This has implications for RQ2, as it shows that Media Topics, while not without its problems, can still be useful for enhancing discoverability. Although there is some overlap in terms, this could be overcome by discussing with colleagues when there is uncertainty and deciding on the terms which will best enhance discoverability for users. This is also in line with Ob2 of my research, as outlined in section 1.2. In a professional context, it would be natural to consult with colleagues and utilise their expertise and judgement in cases of ambiguity.

4.1.2.2. Alternative controlled vocabularies

Media Topics does not seem well-suited for certain domains. As my research is exploratory, this provides an opportunity to explore if other controlled vocabularies would be more applicable to my data. Figure 13 below shows terms taken from LCSH and the UK Archival Thesaurus (UKAT), alongside terms from Media Topics. For comparison purposes, I searched for terms for concepts that were not well described by Media Topics, as identified in section 4.1.1.2.

	MEDIA TOPICS	**LCSH**	**UKAT**
WAR	War	World War, 1914-1918	First World War (1914-1918)
SCOTTISH INDEPENDENCE	Referenda	Autonomy and independence movements; National independence movements [variant]	Independence movements; sovereignty; referenda
FOOTBALL	Soccer	Soccer; Association football [variant]; Football (Soccer) [variant]	Association football

Figure 13: Comparison of controlled vocabularies

The terms found in both LCSH and UKAT seem to be more descriptive than Media Topics. This is particularly true for terms relating to war, as both LCSH and UKAT specify the name and date of the war being described. For Scottish independence, the terms offered by LCSH and UKAT also seem more useful, as they include concepts of sovereignty and autonomy. In contrast, ‘referenda’ only encompasses the act of voting on a proposal.

However, like Media Topics, LCSH uses ‘soccer’ rather than ‘football’. This is not surprising, as it has been noted that LCSH tends to favour a North American perspective (Ragaller and Rafferty, 2012). It does suggest ‘Football (Soccer)’ as a variant term. This could offer a compromise between literary and user warrant in that it might enhance discoverability for a wider audience, whilst remaining true to the content it is describing. However, the variant terms are not the preferred terms. In addition, while LCSH is widely used, it has faced criticism for its complexity (Walsh, 2011).

UKAT was created in recognition of the fact that established controlled vocabularies sometimes do not serve the needs of the UK archival community (UKAT, 2019). The term for football – 'Association football' – sounds quite formal, but does seem to be more consistent with common usage within the UK.

UKAT is a thesaurus rather than a taxonomy like Media Topics. This makes it a more structurally complex type of controlled vocabulary, in which relationships between terms are clearly shown (National Information Standards Organization, 2010). In practice, this might be easier to use in a professional setting as it signposts the metadata creator to the correct or preferred term to use.

Regarding RQ2, the assignation of subject terms is an important part of enhancing discoverability of content in a subtitle archive. Therefore, the choice of controlled vocabulary must be carefully considered. As shown in figure 13 above, some controlled vocabularies offer terms that are more useful or applicable than others. However, they all come with limitations. Media Topics, LCSH and UKAT are only a small sample of the established controlled vocabularies already in existence, and so it would be advisable to explore more options before deciding upon one.

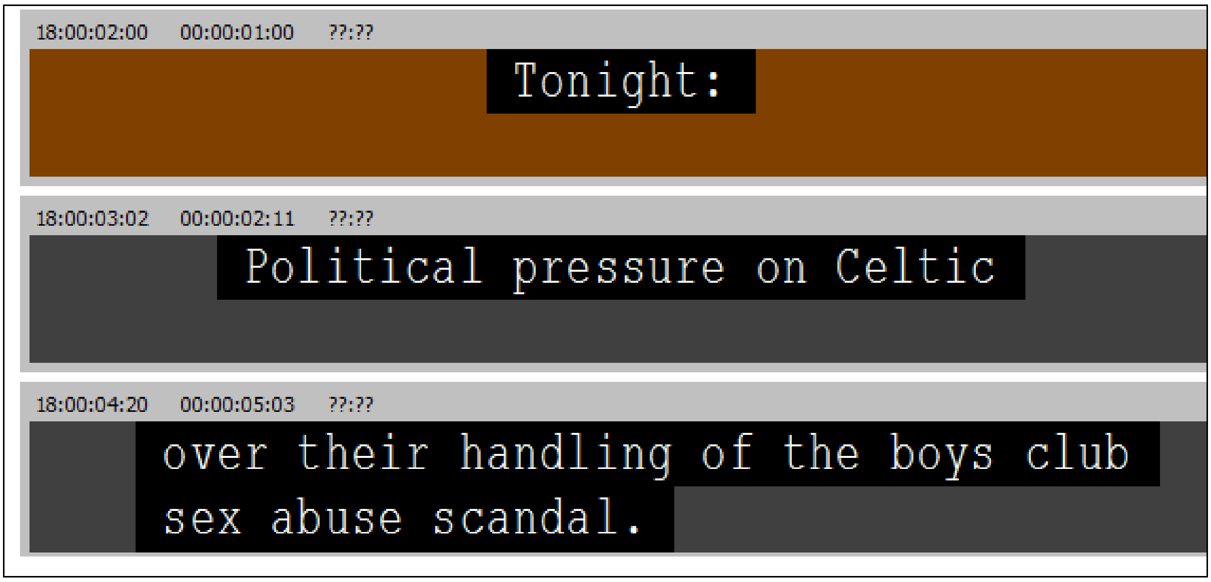
4.2. Free-text keywords

In order to overcome controlled vocabulary limitations, as outlined in 4.1.2.1, I relied heavily on the use of free-text keywords to supplement Media Topics. My initial expectation was that I would use the second subject element iteration selectively, for things like personal names. As news is ostensibly about people, it was not surprising that many personal names were used throughout both datasets.

I entered personal names as values in both the 'subject' element when the person involved seemed to play a prominent role in the aboutness of the item. Additionally, I entered personal names in the 'creator' and 'contributor' elements. In the case of creator, this was always a member of the STV news team, either the news anchor who reads the Ulays or the reporter who is responsible for creating, compiling and delivering the news package.

The only items where the creator was ambiguous were clips. As outlined in section 3.2, clips are short soundbites. The person providing the soundbite could be a politician, sportsman or woman, or member of the public. These people are not the creators of the clip. Rather, if their names are known, they would be listed in the 'contributor' element. The creators of the clips are unknown, as it is not made clear which member of the STV news team is responsible for editing and uploading these. Therefore, I often left the creator element blank.

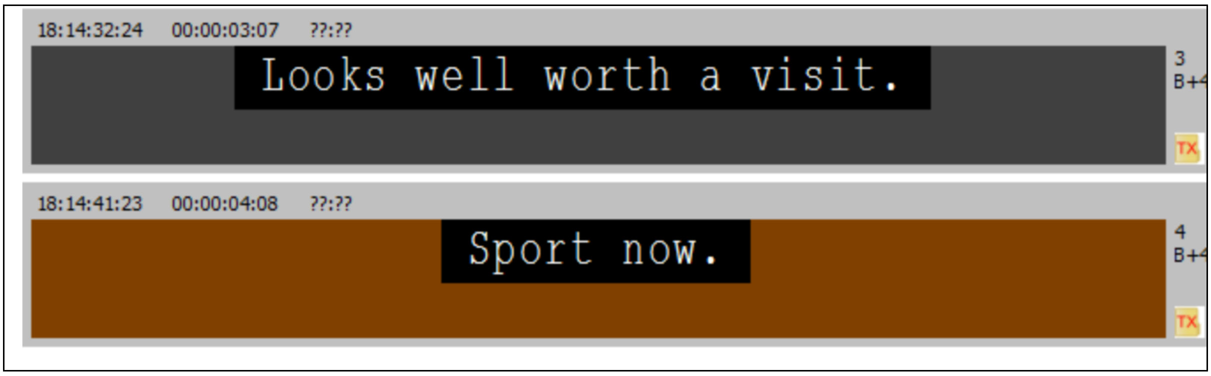
Further instances that required free-text keywords were in items that were interstitial or promotional in nature. By interstitial, I am referring to items that serve a different function to more typical news content. However, in accordance with my methodology, as set out in section 3.7, I treated everything in the running order as a discrete item. Figures 14 and 15 show typical examples of interstitial material. Figure 14 shows an extract of the item G-OPENERS. This is the first item that is transmitted at the beginning of the news programme, in which the news anchors introduce the upcoming stories.



```
18:00:02:00    00:00:01:00    ??:??
Tonight:

18:00:03:02    00:00:02:11    ??:??
Political pressure on Celtic

18:00:04:20    00:00:05:03    ??:??
over their handling of the boys club
sex abuse scandal.
```

Figure 14: Extract of G-OPENERS

Figure 15, G-SPORTHAND, acts as a transition point between the main news coverage and the sports coverage. During this section, the news anchor might exchange ad-libbed words with the sports presenter.



18:14:32:24 00:00:03:07 ??:??

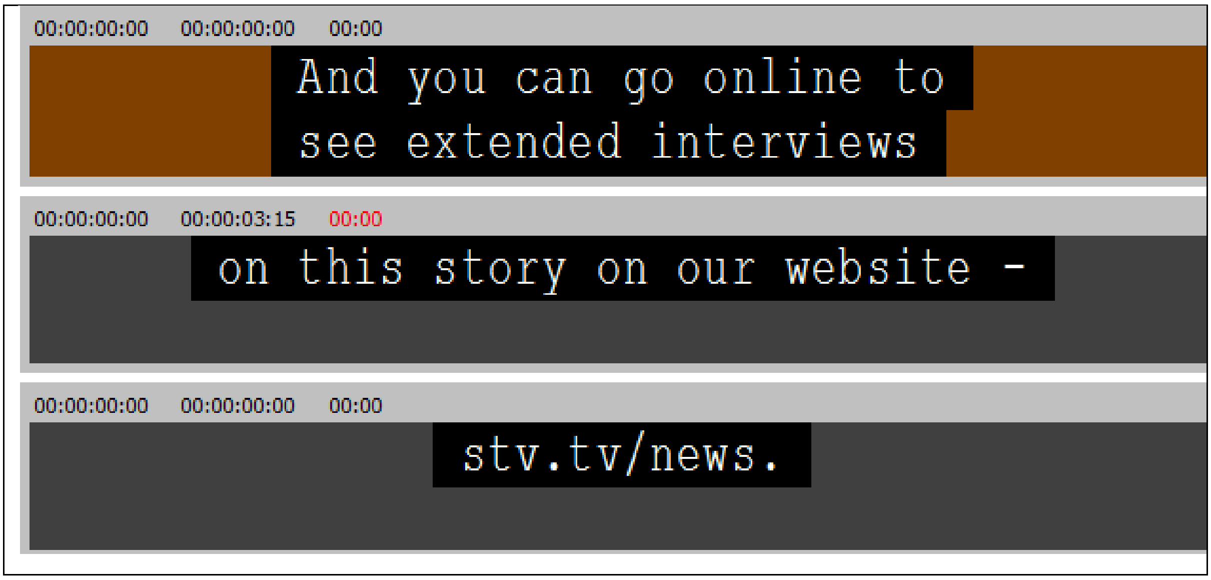
Looks well worth a visit.

18:14:41:23 00:00:04:08 ??:??

Sport now.

Figure 15: G-SPORTSHAND interstitial material

There are also several promotional items, including publicity for a now-defunct digital channel called STV Glasgow, a mention of the STV news website, a quick look at the upcoming weather forecast, and a reminder to viewers to tune in to STV's current affairs programme, Scotland Tonight. Figure 16 shows the promotional item for the STV news website.



00:00:00:00 00:00:00:00 00:00

And you can go online to
see extended interviews

00:00:00:00 00:00:03:15 00:00

on this story on our website -

00:00:00:00 00:00:00:00 00:00

stv.tv/news.

Figure 16: G-WEBSITE promotional material

4.2.1. Analysis of free-text keywords

4.2.1.1. Personal names

As discussed in section 3.6, I did not use LCNAF or LCSH for my personal name entries as I did not think this would be suitable for my datasets. Having conducted the annotation, I feel that my initial expectation was correct in this regard. Of the 18 individually named reporters across both datasets, only one had an entry in LCNAF. In terms of personal names as subject terms, these were also not well represented by LCSH. As an example, I searched LCSH for all the names that occurred across both datasets that had some connection to sport. Figures 17 and 18 below highlight how few of the names appeared as entries in LCSH.

NAME	ENTRY IN LCSH
NEIL LENNON	Lennon, Neil
PETER LAWWELL	✗
JOHN KENNEDY	✗
DAMIEN DUFF	Duff, Damien, 1979-
STEVIE WOODS	✗
SHELLEY KERR	✗
KYLIE COCKBURN	✗
STEVIE RAY	✗
LEONARDO SANTOS	✗

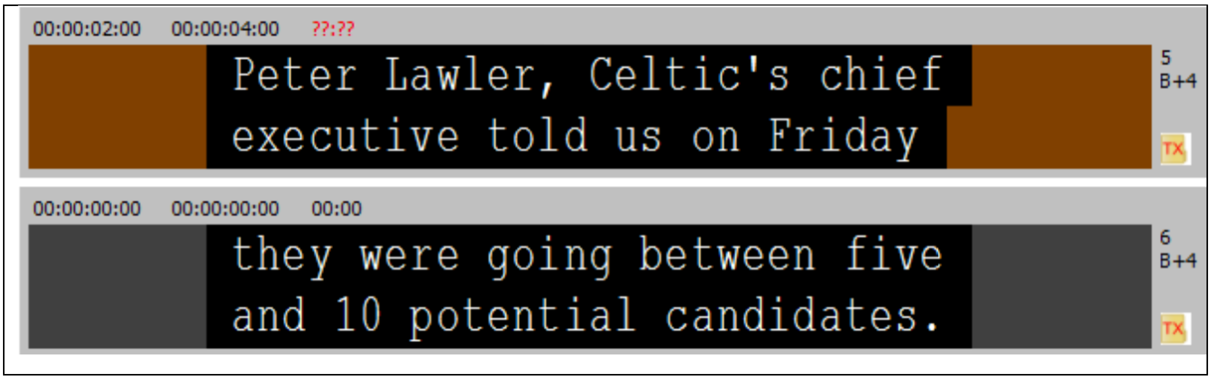
Figure 17: Sports personalities appearing in 2019 dataset

NAME	ENTRY IN LCSH
ROY KEANE	Keane, Roy, 1971-
STEVE CLARKE	✗
MARTIN O'NEILL	O'Neill, Martin, 1952-
OWEN COYLE	Coyle, Owen, 1966-
MALKY MACKAY	✗
ANDY MURRAY	Murray, Andy, 1987-
FERNANDO VERDASCO	✗
AL KELLOCK	✗
JAMIE HAMILL	✗
ANN BUDGE	✗
GARY HOLT	✗
NEIL ADAMS	Adams, Neil, 1965-
STEPHEN GALLACHER	✗
THONGCHAI JAIDEE	✗

Figure 18: Sports personalities appearing in 2014 dataset

It is not surprising that so few entries were found in LCSH, as the type of names that appear in STV news may not be well known outside of Scotland. Additionally, the people being referred to are often up-and-coming in their fields, as in the case of young footballers. Consequently, they would be unlikely to have LCSH entries. It therefore seems more practical to enter the names as they appear, using free-text keywords.

However, one problem did arise which, while not commonplace, demonstrates why authority control over named entities is useful. In the 2014 dataset, I encountered a name that had been misspelled as Peter Lawler instead of Peter Lawwell. Figure 19 shows this error in the original subtitle files. I only noticed this error because the name Peter Lawwell is regularly mentioned on STV news bulletins. Had I not noticed the error, then the incorrect name Peter Lawler could have been mistakenly entered into the metadata.


00:00:02:00 00:00:04:00 ??:??

Peter Lawler, Celtic's chief
executive told us on Friday

5
B+4

TX

00:00:00:00 00:00:00:00 00:00

they were going between five
and 10 potential candidates.

6
B+4

TX

Figure 19: Peter Lawwell name misspelling

In my personal experience, these kinds of misspellings are not commonplace. However, as discussed in section 2.6.1, the combination of working in live, pressurised conditions and using voice recognition software can cause errors to occur. What this draws attention to is the larger problem related to our working practices, and the conditions under which subtitles are produced. In response to RQ2, misspellings and other errors could limit discoverability of content. Therefore, the person responsible for creating the metadata records would have to ensure any errors in the subtitles are not transferred to the metadata. This also relates to RQ4, as this kind of attention to detail inevitably has a trade-off in terms of time taken to thoroughly check all the content and ensure it is correct. As discussed above, I noticed the error because I was familiar with the name Peter Lawwell.

However, there is no guarantee that all similar errors will be picked up on and thus, the trade-off may be that sometimes names will not be spelled correctly if there is no authority file to ensure it is entered in a consistent manner.

4.2.1.2. Interstitial and promotional material

Conducting a subject analysis on interstitial material proved challenging because they are not really 'about' anything. A typical news item has relevant themes and concepts to extract. In contrast, interstitial and promotional material, semantically speaking, has very little to analyse. As a result, I was uncertain which subject terms to apply. For G-OPENERS, I applied both controlled and uncontrolled subject terms, to describe the stories that were being announced. The metadata record for this item is shown in figure 20. However, by applying these subject terms, there is an impression that the content contained within the item itself encompasses all of these topics. In reality, that is not the case. Rather, the item is simply promoting the fact that these stories are coming up later in the programme.

1

TITLE	G-OPENERS
IDENTIFIER	NEWSATSIXGLA020614G-OPENERS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	John MacKay introduces the news headlines
SUBJECT	Referenda; motor car racing; road accident and incident; soccer; war
SUBJECT	Scottish independence; Celtic Football Club; Roy Keane; Steve Clarke; WW1; First World War; Great War
CREATOR	John MacKay
CREATOR ROLE	Commentator

Figure 20: Metadata record for G-OPENERS

Promotional items, too, had a lack of suitable controlled vocabulary terms. Media Topics only offers 'printing/promotional service', which is this is intended for the description of printed material, rather than promotions for broadcast news content. I accommodated for the lack of controlled vocabulary by using the keyword ‘promo’ to indicate the nature of the item.

These interstitial and promotional items might seem comparatively unimportant alongside other items, which have a definite subject matter. However, in accordance with Balbi’s (2011) all-in approach to preservation, as discussed in section 2.3.1, every item was considered equal in terms of its value. Allen and Schalow (1999) argue that there is research interest in the physical layout of newspapers. Therefore, it stands to reason that there may also be interest in the layout of broadcast news, which includes the interstitial and promotional items found throughout each of my datasets.

Regarding RQ2, the way in which I treated interstitial material does not enhance discoverability. Rather, it gives the false impression that there is relevant content within the item, when in fact, the item only indicates that this content exists in other resources. This is less true for promotional items as they do contain content. However, a lack of suitable subject terms mean they are not well described. I would argue that, in a professional setting, interstitial and promotional items should be treated differently from the main news content. They serve a different, but equally valid, function. However, having conducted my annotation, I realise that my current metadata schema does not provide a suitable way to indicate the different function that these items serve.

4.2.3. Suggestions to improve free-text keywords

I was unable to apply keywords in a consistent and intuitive manner, largely because I had no policy or guidelines to inform me in this process. However, as my research is exploratory, this provides an opportunity to develop best practice guidelines based on the difficulties I encountered. The sections below offer some suggestions for improvement. Section 4.2.3.1 considers creating a specialised set of terms to accommodate the needs of my datasets, while section 4.2.3.2 considers the strengths and limitations of outsourcing the assignation of free-text keywords to external users.

4.2.3.1. Create a small-scale controlled vocabulary

Creating a controlled vocabulary is a significant undertaking in terms of establishment and maintenance. It is therefore generally not advised for large collections when there is likely to be overlap with existing controlled vocabularies (National Information Standards Organization, 2010; Suominen and Mader, 2013). However, this option becomes more feasible when the set list of terms is very small and localised (Miller, 2011), as in the case with the interstitial and promotional items I identified throughout both of my datasets.

Figure 21 shows the interstitial material, alongside a proposed controlled vocabulary term and a scope note indicating how the term should be applied. A similar set of terms is shown in figure 22 for use with promotional items.

ITEM NAME	CONTROLLED VOCABULARY TERM	SCOPE NOTE
OPENER	Headlines	Use when news anchor(s) announce upcoming stories
SPORTHAND	Handover	Use when there is a transitional segment from general news to sports news
BACK	Handback	Use when sports presenter closes the segment
BYE	Closer	Use when news anchor(s) close the programme

Figure 21: Controlled vocabulary terms for interstitial material

ITEM NAME	CONTROLLED VOCABULARY TERM	SCOPE NOTE
TONIGHT	Scotland Tonight (promo)	Use when news anchor(s) mention Scotland Tonight
WEBSITE	Website (promo)	Use when viewers are directed to STV news website
CITY	Digital channel (promo)	Use when reference is made to STV digital channel(s)

Figure 22: Controlled vocabulary terms for promotional material

Rather than applying this controlled vocabulary in the subject element, another option would be to introduce a new element to describe the content type. XMLNews uses the element ‘fixture’ to denote features that occur on a regular basis (XMLNews-Meta Technical Specification, 1999). Interstitial and promotional material features regularly in STV news, which makes the ‘fixture’ element a suitable choice.

It would also be relatively straightforward to create name authority files for all STV news presenters and reporters. As mentioned in section 4.2.1.1, there are only 18 named STV news staff throughout both datasets. Thus, in terms of RQ4, there would be an initial time investment in terms of establishing the name authority files. However, this would save time overall when adding these names into the metadata records. I would not recommend creating authority files for other personal names, such as sports personalities or any other named people in the datasets. The time it would take to create these files would be far greater than the time it would take to create 18 name authority files for STV news staff. It remains more practical to simply enter these names as free-text keywords where relevant.

This solution of a small-scale controlled vocabulary has relevance to RQ2 of this dissertation. With regards to interstitial and promotional material, existing controlled vocabularies do not meet the needs of my proposed subtitle archive. Equally, keywords can be applied in an inconsistent manner, which affects the quality of metadata. Therefore, creating a small list of controlled vocabulary terms could result in a richer level of subject indexing, thus enhancing discoverability of content.

4.2.3.2. Collaborative tagging

Collaborative tagging would allow external users could assign keywords to items, thus providing another dimension to the subject indexing process. Disadvantages of user-assigned tags include a lack of vocabulary control, and the subjectivity involved in choosing terms (Joudrey, Taylor and Miller, 2015). This calls attention to the need for guidelines to ensure consistency and a degree of standardisation across the resources (Guy and Tonkin, 2006; Haynes, 2018).

However, the process of subject analysis can be subjective (Library of Congress, 2016). As discussed in section 4.1.2.1, intercoder reliability testing shows that controlled vocabulary terms with similar meanings can be assigned by independent coders, due to limitations of the controlled vocabulary (Suominen and Mader, 2013). Therefore, subjectivity does not necessarily negate the usefulness of collaborative tagging. For a fuller discussion on the advantages and disadvantages of collaborative tagging, see section 2.1.2.2.

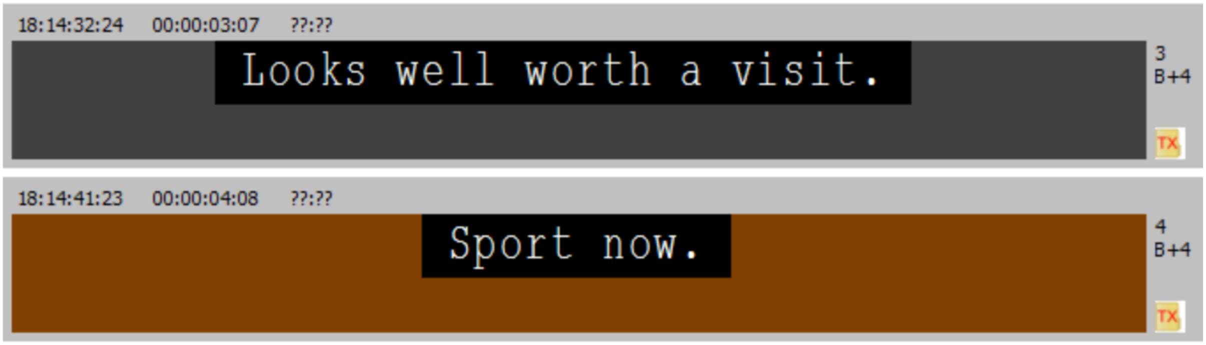
Neal (2008) has argued that controlled vocabularies and collaborative tagging do not have to conflict with each other and that combining them may be advantageous. It should be noted that this study differs from mine in that it concerns news photographs. Photographs are more obvious candidates for collaborative tagging, as they typically do not contain any text from which to extract meaning (Haynes, 2018). In contrast, subtitles only contain text. However, a key point of relevance here is that controlled vocabularies "are not updated quickly enough to remain relevant to current news events" (Neal, 2008, p. 211). As discussed in section 4.2.1.1, one reason for using uncontrolled personal names is because the people are still emerging in their fields and do not have an authorised term. This idea could be extended to encompass not just names, but new and emerging concepts more widely, of the kind seen in broadcast news. Therefore, combing controlled terms with collaborative tagging could ease the pressure on the metadata creators and allow users to add their own potentially valuable insights.

Regarding RQ2, a collaborative tagging system could enhance discoverability if it was implemented with appropriate guidelines to ensure users created high quality metadata. This has important implications for RQ3. The subtitle archive dates back to 2009. In order to retroactively annotate the vast amount of records from the past, willing volunteers would need to be enlisted to help with the indexing process. Without some form of external help, the project would be unfeasible. Following on from both these points, in relation to RQ4, there is inevitably a trade-off. Collaborative tagging may result in tags which are not of a high standard. However, without this external assistance, the quality of the metadata created in-house would, by necessity, be less detailed due to the time needed to annotate such a large volume of resources.

4.3. Item-level metadata

As explained in methodology section 3.7, I treated each individual element of the news running order as a discrete item. However, there are natural linkages between some of these items. This is exemplified by figures 24 and 25 on the following page. Figure 24 shows a Ulay for an item about mixed martial arts, called G1-UFC. Figure 25 is the associated clip that was broadcast directly after the Ulay, called G2-UFC. If G2-UFC was viewed in isolation, it would be lacking contextual information which is provided by G1-UFC. For example, the information shown in figure 24 tells viewers the overall topic of the item, and also gives an indication of who the speaker is in figure 25. This represents a whole-part relationship, in which one item is a component which is logically related to another item (Miller, 2011). These relationships were found frequently throughout both datasets. I used the elements 'Has Part' and 'Is Part Of' to indicate the linkages between items.

Items can also be linked through the presenters' use of ad-libbed sentences. As explained in section 3.2, presenters will often intersperse offhand, spontaneous comments throughout the programme. These comments are sometimes directed to other presenters, and sometimes they are directed to the viewers at home. An example of this is shown in figure 23. The comment "Looks well worth a visit" refers to the item that preceded it, called G2-HILL.


18:14:32:24 00:00:03:07 ??:??

Looks well worth a visit.

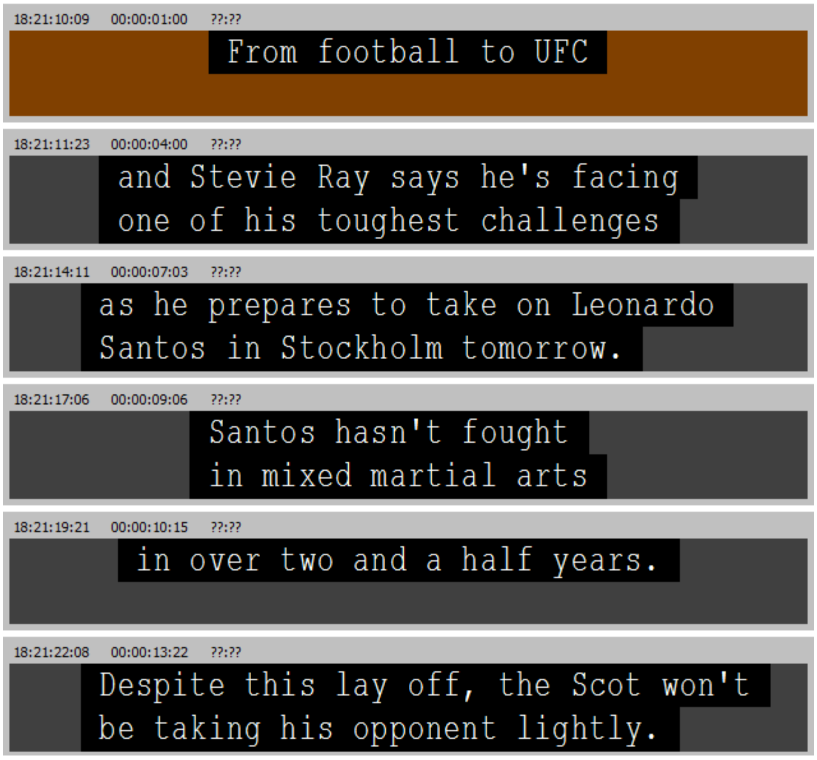
3
B+4
TX

18:14:41:23 00:00:04:08 ??:??

Sport now.

4
B+4
TX

Figure 23: Sport handover ad-lib



18:21:10:09 00:00:01:00 ??:??
From football to UFC

18:21:11:23 00:00:04:00 ??:??
and Stevie Ray says he's facing
one of his toughest challenges

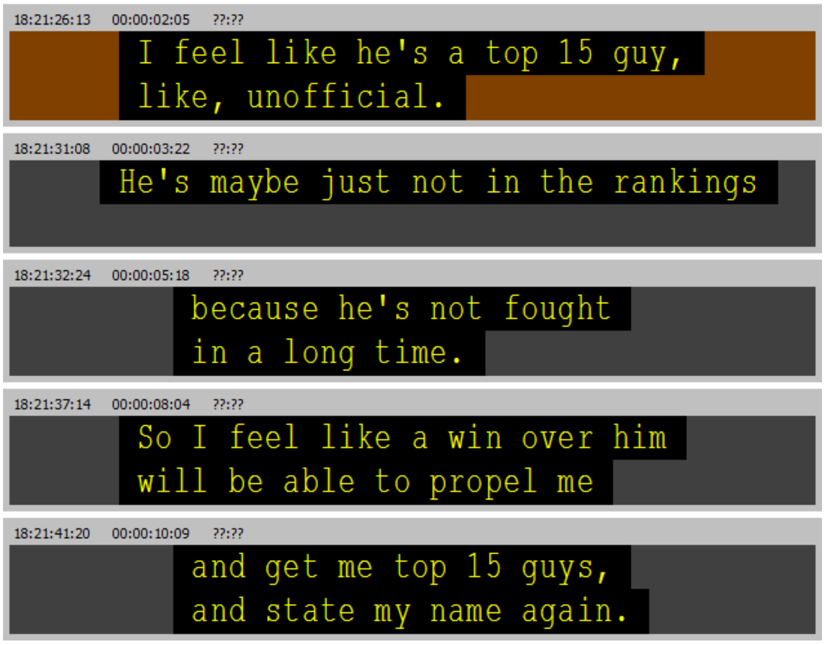
18:21:14:11 00:00:07:03 ??:??
as he prepares to take on Leonardo
Santos in Stockholm tomorrow.

18:21:17:06 00:00:09:06 ??:??
Santos hasn't fought
in mixed martial arts

18:21:19:21 00:00:10:15 ??:??
in over two and a half years.

18:21:22:08 00:00:13:22 ??:??
Despite this lay off, the Scot won't
be taking his opponent lightly.

Figure 24: Ulay for G1-UFC



18:21:26:13 00:00:02:05 ??:??
I feel like he's a top 15 guy,
like, unofficial.

18:21:31:08 00:00:03:22 ??:??
He's maybe just not in the rankings

18:21:32:24 00:00:05:18 ??:??
because he's not fought
in a long time.

18:21:37:14 00:00:08:04 ??:??
So I feel like a win over him
will be able to propel me

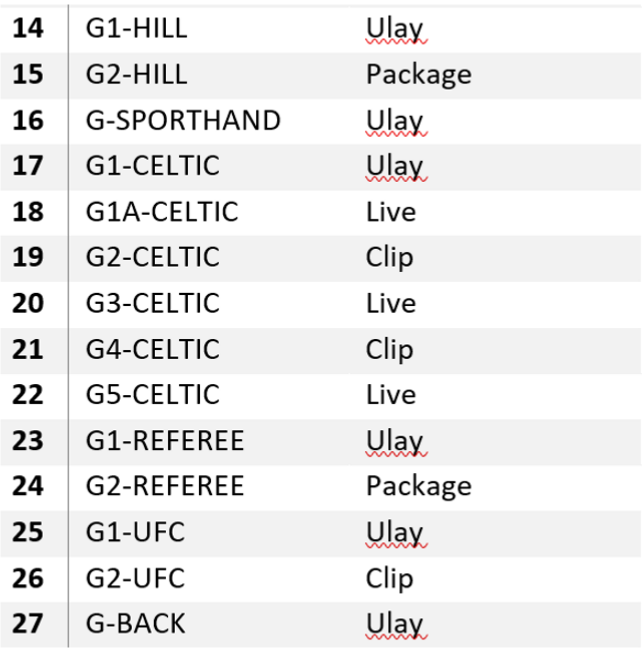
18:21:41:20 00:00:10:09 ??:??
and get me top 15 guys,
and state my name again.

Figure 25: Ulay for G2-UFC

4.3.1. Analysis of item-level metadata

4.3.1.1. Whole-part relationships

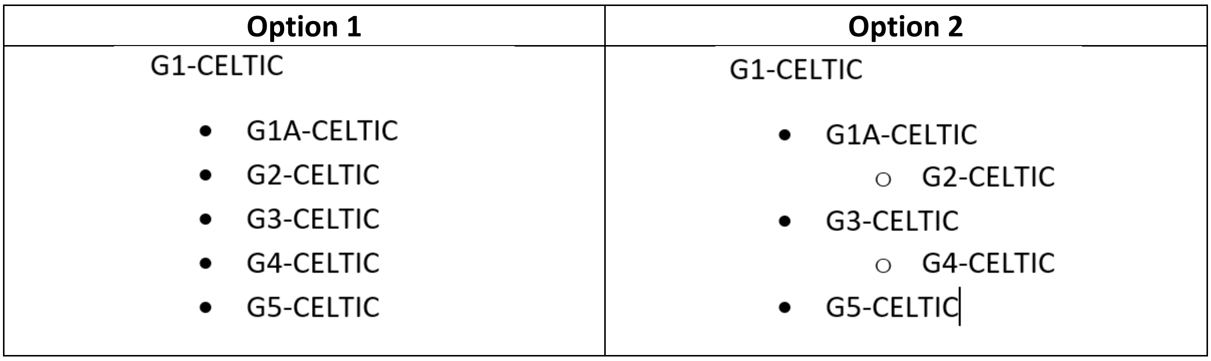
My initial expectation was that most items in the running order would have a straightforward relationship, like the examples shown in figures 24 and 25. However, other items had relationships that were much more complex to interpret. For example, figure 26 is an extract of the news running order from the 2019 dataset. Items 17-22 all have the same title (CELTIC) and the numbers indicate they are arranged in sequential order. Each individual item contains a slightly different component of the story. The entire CELTIC segment consists of a sports presenter talking in the studio (Ulay), a reporter on location (Live), and soundbites of the football management team (Clip).



14	G1-HILL	Ulay
15	G2-HILL	Package
16	G-SPORTHAND	Ulay
17	G1-CELTIC	Ulay
18	G1A-CELTIC	Live
19	G2-CELTIC	Clip
20	G3-CELTIC	Live
21	G4-CELTIC	Clip
22	G5-CELTIC	Live
23	G1-REFEREE	Ulay
24	G2-REFEREE	Package
25	G1-UFC	Ulay
26	G2-UFC	Clip
27	G-BACK	Ulay

Figure 26: Extract of news running order

When creating metadata records for each of these items, the process of indicating the linkages between them became rather convoluted. Although DCMI has a flat structure, I conceptualised the linkages in a hierarchical structure, in order to determine which items belonged in the 'Has Part' or 'Is Part Of' elements. Figure 27 below shows two different possibilities for conceptualising the linkages between the CELTIC items.



Option 1	Option 2
G1-CELTIC • G1A-CELTIC • G2-CELTIC • G3-CELTIC • G4-CELTIC • G5-CELTIC	G1-CELTIC • G1A-CELTIC ○ G2-CELTIC • G3-CELTIC ○ G4-CELTIC • G5-CELTIC

Figure 27: Conceptualised linkages between items

Option 1 follows the same pattern as shown in figures 24 and 25. That is, G1-CELTIC is a Ulay and every item that falls underneath G1-CELTIC is logically related, and therefore indicated by the 'Is Part Of' element. However, option 2 shows how the items could be further nested within each other. There is an argument for the structure shown in option 2, because some of the items seem to logically depend on others. For example, G2-CELTIC is a clip, and the content contained within that clip only makes logical sense when viewed directly after G1A-CELTIC.

I followed the conceptual structure shown in option 2 when describing the linkages between the items. In addition, I included G-SPORTHAND (item number 16, shown in figure 26) as a superordinate from which all the items below fall under. As discussed in section 4.2.3.1, G-SPORTHAND represents interstitial material, and I would argue that it should be treated differently from the other items. However, at the time I was conducting the annotation, I treated G-SPORTHAND as though it was in a superordinate position because this marked the introduction of the sports section of the programme.

This demonstrates that there are various interpretations possible regarding the linkages between items. This may mean that there is a lack of consistency in how metadata is entered in the metadata record for each item. In turn, this has implications for RQ2 as, if these relationships are not made explicit, users will be unable to locate relevant, sometimes even necessary, contextual information.

4.3.1.2. Ad-libbed sentences

Figure 23 shows an ad-libbed sentence made by the news presenter about the preceding item, G2-HILL. The item in which the ad-lib occurs is G-SPORTHAND, which introduced the start of the sports coverage. Thus, there is a reference to a different item which is not represented by the metadata record for G-SPORTHAND (Miller, 2011). However, unlike the linkages discussed in section 4.3.1.1, neither G-SPORTHAND or G2-HILL are a logical part of the other.

The ad-libbed comment does not concern sport, or introduce any following items. It is simply a spontaneous comment made about another resource, referring to the reopening of a visitor attraction called Hill House. This connection between items cannot be expressed by the 'Is Part Of' element because G-SPORTHAND is not a part of G2-HILL. They do not have the same themes or story content throughout.

My metadata schema, as set out in section 3.6, had no way of accommodating these types of linkages. Consequently, users viewing the content contained in G-SPORTHAND would not be provided with any context for the comment "Looks well worth a visit". With regards to RQ2, showing these linkages is useful, as they point users to other items that may be relevant to their interests.

4.3.2. Suggestions to improve item-level metadata

The DCMI 'Relation' element offers various qualifiers to indicate different types of relationships between items (Miller, 2011). I only included two of these qualifiers in my metadata schema – 'Has Part' and 'Is Part Of'. The following suggestions offer alternative ways of indicating connections between items. The first suggestion, set out in section 4.3.2.1, considers the addition of a further element, while the suggestion outlined in 4.3.2.2 discusses why it may be beneficial to adopt a different approach to the granularity of description.

4.3.2.1. References element

In addition to the 'Has Part' and 'Is Part Of', DCMI offers 'References' as a qualifier to indicate relationships between resources (Hillmann, 2005). The 'References' qualifier is used to indicate “a related resource that is referenced, cited, or otherwise pointed to be the described resource” (Miller, 2011, p. 121). As such, it would be ideally suited to ad-libbed sentences like the one shown in figure 23 in section 4.3. This connection can be reciprocal, with the use of a corresponding ‘Is Referenced By’ qualifier. Figures 28 and 29 show what the metadata records for G-SPORTHAND and G2-HILL would look like with these additional elements.

In terms of RQ2, users would more easily be able to locate resources that have some relevance to the item described by the metadata record. Thus, a simple addition of the ‘References’ element would enhance discoverability and ensure these linkages are accounted for.

16

TITLE	G-SPORTHAND
IDENTIFIER	NEWSATSIXCEN310519G-SPORTHAND
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Gordon hands over to Raman for the sport
SUBJECT	Sport
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G1-CELTIC
HASPART	G1A-CELTIC
HASPART	G2-CELTIC
HASPART	G3-CELTIC
HASPART	G4-CELTIC
HASPART	G5-CELTIC
HASPART	G1-REFEREE
HASPART	G2-REFEREE
HASPART	G1-UFC
HASPART	G2-UFC
HASPART	G-BACK
REFERENCES	G2-HILL

Figure 28: Metadata record for G-SPORTHAND with additional ‘References’ element

15	
TITLE	G2-HILL
IDENTIFIER	NEWSATSIXCEN310519G2-HILL
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Conservationists are preparing to re-open Hill House in Helensburgh following extensive work to preserve it
SUBJECT	Monument and heritage site; renovation
SUBJECT	Conservation; Hill House; Charles Rennie Mackintosh; Helensburgh
CREATOR	Susan Ripoll
CREATOR ROLE	Reporter
ISPARTOF	G1-HILL
ISREFERENCEDBY	G-SPORTHAND

Figure 29: Metadata record for G2-HILL with additional 'Is Referenced By' element

4.3.2.2. Collection-level items

Rather than creating an individual metadata record for every item in the dataset, it would be more practical, in terms of time and linkages, to consider related items as a collection. This refers to the granularity of description. Throughout the annotation process, I was creating metadata at a very high level of granularity. In contrast, a collection-level metadata record is less granular, as it describes the entire contents contained within the collection (Miller, 2011). As discussed in section 4.1.1.1, I found duplication of subject terms between individual metadata records. This was further confirmed in section 4.1.2.1, when I asked my colleague Alex to assign subject terms to Ulay items and their associated package items. He, too, tended towards duplication of subject terms between individual items. Thus, although both Alex and I were creating metadata at the item-level, these items were nearly indistinguishable in terms of overall subject matter. This suggests that it would be more advantageous to create metadata records at the collection-level than the item-level.

Regarding RQ2, users would be able to obtain all relevant content if resources are described at the collection-level. This also eases pressure on the metadata creator, as less time will be needed to ensure all relevant linkages between items are entered into the metadata record. Thus, in terms of RQ4, the trade-off in terms of time investment would be significantly reduced by creating collection-level metadata records.

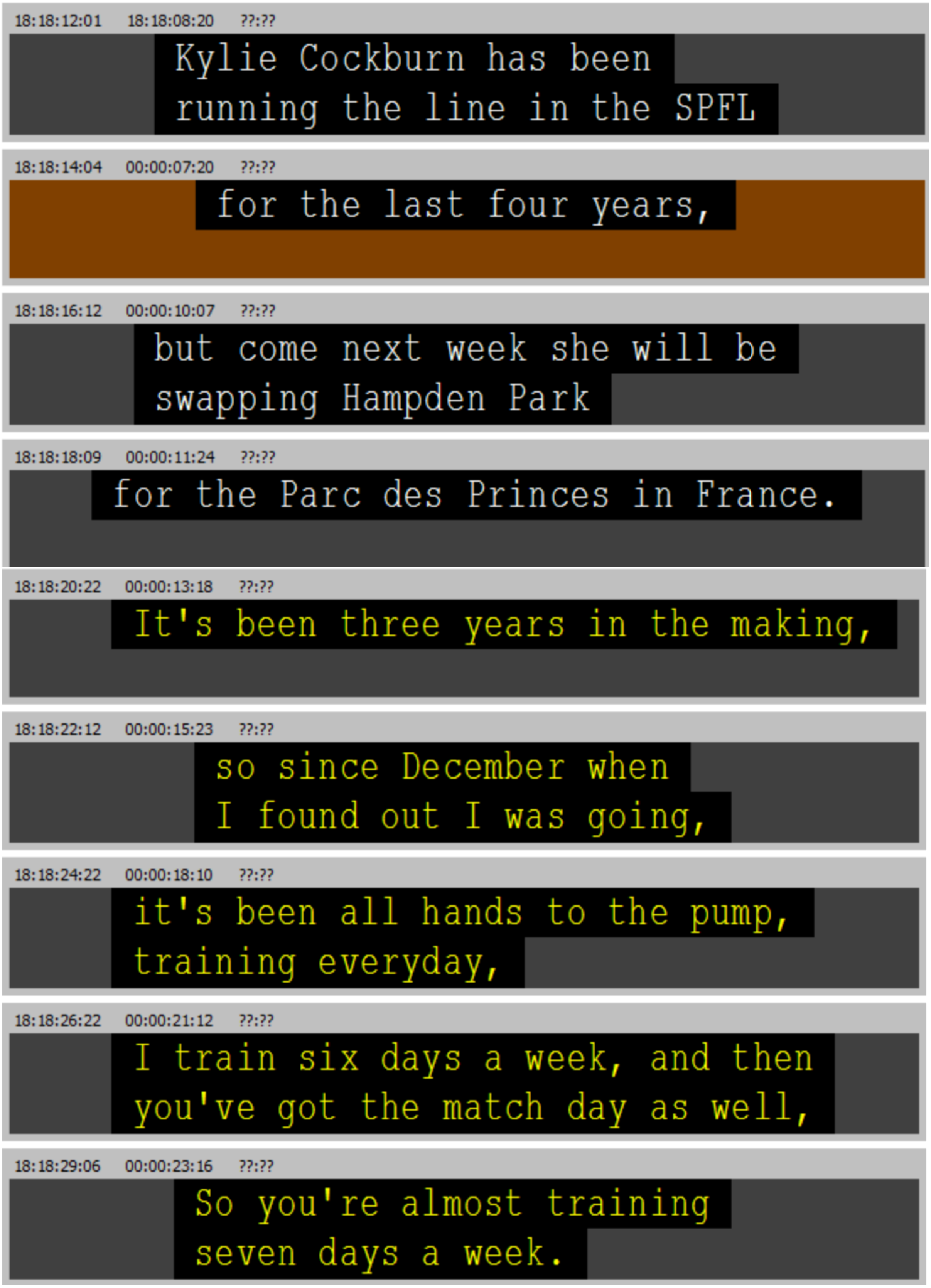
4.4 Speaker identification

As first discussed in the introduction to this dissertation, Snider (2000) proposed an archive of subtitles could serve as a valuable alternative means of obtaining content from otherwise-inaccessible broadcast archives. Throughout this analysis chapter, I have identified areas that require further exploration in order to ensure that a subtitle archive is accessible and useful to external users. However, one final that I will discuss was also alluded to by Snider, which relates to the issue of speaker identification.

Subtitles represent the speech of various people. In my personal experience at STV, name tags are rarely used to denote speakers, particularly in news programmes. This is for two reasons. Firstly, different coloured text indicates when a change of speaker takes place. Secondly, the subtitles are typically displayed on screen at the same time as the speaker. Thus, I was able to identify speakers in the 2019 dataset because I had access to the video footage. In contrast, I was unable to identify speakers in the 2014 dataset unless this information was otherwise made explicit in the text. Therefore, where possible, I entered the names of known speakers in the 'Contributor' element.

4.4.1. Analysis of speaker identification

As this research is at an exploratory stage, much of the focus has been on creating metadata to turn the existing dark archive into something that is more accessible. Thus, there has been little consideration paid to what the end product of an accessible subtitle archive might look like. However, one possibility is that the user would be presented with a text document, which may or may not retain the different coloured text, used to denote a change of speaker. When there are only one or two speakers, reasonable inferences can be made about which segments of text can be assigned to which speaker.



18:18:12:01 18:18:08:20 ??:??

Kylie Cockburn has been
running the line in the SPFL

18:18:14:04 00:00:07:20 ??:??

for the last four years,

18:18:16:12 00:00:10:07 ??:??

but come next week she will be
swapping Hampden Park

18:18:18:09 00:00:11:24 ??:??

for the Parc des Princes in France.

18:18:20:22 00:00:13:18 ??:??

It's been three years in the making,

18:18:22:12 00:00:15:23 ??:??

so since December when
I found out I was going,

18:18:24:22 00:00:18:10 ??:??

it's been all hands to the pump,
training everyday,

18:18:26:22 00:00:21:12 ??:??

I train six days a week, and then
you've got the match day as well,

18:18:29:06 00:00:23:16 ??:??

So you're almost training
seven days a week.

Figure 30: Extract of G2-REFEREE

Figure 30 shows an extract of an item which features two speakers. The first speaker, in white text, is the reporter, whose name I entered into the ‘Creator’ element. The second speaker, in yellow text, is Kylie Cockburn, who is being interviewed by the reporter about her experience of being a football referee. In this example, the colours and the two names entered into the metadata record make it possible to infer which segments of text can be assigned to each speaker.

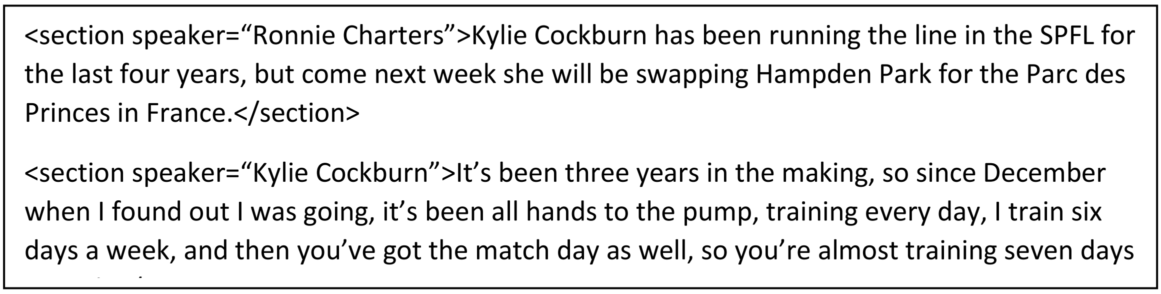
However, this becomes more complicated when multiple speakers are introduced. There may be clues in the text, or colours to indicate a change of speaker, but oftentimes the speaker will be ambiguous. This has implications for RQ2 as, without explicit indications as to speaker identification, users are left to make inferences about who the words can be attributed to. As the same time, this also impacts on RQ3. It is not possible to consult the video footage for historic data, which further limits the possibility of entering the names of speakers into the metadata records.

4.4.2. Suggestions to improve speaker identification

The following suggestion will consider the use of markup to enter metadata directly into the content itself. For a fuller discussion of markup, see section 2.7.1.

4.4.2.1. Markup

Tags can be used to indicate features within the text (Riley, 2017). In markup, tags are elements, containing information about the thing that is being described. The attributes have values (XML Tree, 2019). Using figure 30 as an example, I could create an element called <section> to delineate blocks of text. I could then assign the attribute ‘speaker’ to the section element, which would have an associated value. An example of how this would look is shown in figure 31 below.



<section speaker=“Ronnie Charters”>Kylie Cockburn has been running the line in the SPFL for the last four years, but come next week she will be swapping Hampden Park for the Parc des Princes in France.</section>

<section speaker=“Kylie Cockburn”>It’s been three years in the making, so since December when I found out I was going, it’s been all hands to the pump, training every day, I train six days a week, and then you’ve got the match day as well, so you’re almost training seven days

Figure 31: Example of markup

In terms of RQ2, being able to identify speakers within the text would be valuable for users who want to find quotes attributed to particular people. Markup also improves the overall context of the subtitle text, as reduces the need for guesswork on the part of the user. This has implications for RQ4, as there would be a time investment involved in adding this further level of metadata into the content. However, the trade-off would result in more useful and more discoverable content. Regarding RQ3, there will be occasions when speakers cannot be identified because of the lack of associated video footage. This therefore limits the extent to which historic data can be annotated.

5.0. Recommendations and conclusions

The present study was designed to explore what issues arise from improving access to television news subtitles that, after their initial airing, are otherwise left preserved but unutilised in a “dark archive” (Allen and Johnson, 2008, p. 394). The most obvious finding to emerge from the study is that subtitles fall somewhere between traditional newspapers and broadcast news. They are textual in nature but representative of a spoken form of news. This brings together a unique combination of features which cannot easily be accommodated by a traditional metadata schema such as DCMI. Therefore, elements must be drawn from several sources in order to describe the characteristics of subtitles. The analysis suggests that, with further refinements, a subtitle archive could offer something of discernible value to external users.

5.1. Limitations

This is a case study specifically for STV news. Working within the organisation, I was able to gain access to data and insights that would have otherwise been closed off to me. Therefore, these results cannot be easily generalised to other subtitling companies. The two other main subtitle producers within the UK are Red Bee Media and TVT Media (About Us, 2019; We Are TVT Media, n.d.). Unlike the STV subtitling department, which is based in-house, Red Bee Media and TVT Media are external companies. Therefore, their access to subtitling data may be restricted. They also provide services for a variety of clients, rather than only one broadcaster. In contrast, the subtitles at STV are created only for STV-produced content. The other companies may be more likely to encounter complex legal issues around ownership of content, preventing them from carrying out this kind of research.

The generalisability of these results is also subject to certain limitations within STV news. Firstly, I only looked at the 6pm bulletin. This follows a certain format and contains certain types of content that may differ from other STV news programmes. For example, Scotland Tonight is STV’s news and current affairs programme (STV, 2018). It contains extended interviews and more live segments than the 6pm bulletin. This would raise questions about the level of exhaustivity required to index the content, as well as the likely number of errors that result from the increased amount of live subtitling required.

5.2. Research questions

5.2.1. RQ1: What is the value of an archive of television news subtitles?

There are some issues with accessibility of content in newspaper archives but, in general, traditional print news is more widely available than archived broadcast news. At the same time, legal requirements stipulate that subtitles must be produced for broadcast news. However, beyond their one-time usage, they are not being utilised. Therefore, the value of a subtitle archive comes from repurposing them as an alternative means of accessing broadcast news content. For a fuller discussion, see section 2.9.

5.2.2. RQ2: How can we enhance discoverability of content in a subtitle archive?

Common metadata elements associated with discoverability include title, creator and subject (Zeng and Qin, 2008). Of these, title and creator were straightforward to input into the metadata record. Due to my position, I can see the titles assigned to news items. The creators were the news anchors and reporters, all of whom are introduced by name and represented in the subtitles. Thus, information about title and creator is already provided. The area of most concern to me for enhancing discoverability was the subject element, because currently, the subtitle archive has no effective means of searching the content by subject. Instead, subtitle files are preserved in a dark archive.

According to Erickson (2013, n.p.), a dark archive is "typically used for the preservation of content that is accessible elsewhere". In this sense, subtitles function as a stand-in for the original broadcast material. They are representative of the words spoken in the programme. However, it is not true that the original broadcast material is accessible elsewhere. After the original transmission of STV news programmes, viewing possibilities become extremely limited. Some archive footage is available but that either comes at a cost and is used for business to business purposes, or is only available in limited amounts. Herein lies the crux of the argument I have made throughout this dissertation – subtitle files are already preserved in some form, whereas access to broadcast footage is restricted. Therefore, a subtitle archive could provide some useful benefit, but the content must first be made discoverable.

I considered different methods of subject indexing, ranging from uncontrolled keywords to more controlled encoding schemes. As discussed in sections 4.1-4.2, all of these options have strengths and weaknesses. However, they all offer an improvement on the existing subtitle archive, which has no form of resource description. Thus, the application of a small number of terms that describe the items' contents could enhance discoverability. This could be further strengthened with intercoder reliability testing to ensure a level of consistency in how these terms are applied.

I encountered various challenges with creating metadata for my data samples, and a significant section of my analysis in chapter 4 is devoted to issues of subject indexing. This study has identified things that did not work well, and put forward some considerations for improving on these initial attempts. This offers valuable insights into how to enhance discoverability.

5.2.3. RQ3: To what extent can we retroactively annotate records from the past to provide added value?

In traditional print media, historical news content is readily available, and work is continuously underway to improve access to these resources (King, 2005; McKernan, 2014). In contrast, television news is less accessible, in large part because broadcasters are under no obligation to provide access to their content and to do so would place a significant burden on them in terms of time and resources (Kramp, 2014; Cigognetti, 2001).

Ostensibly, STV news subtitle files are preserved. However, in practical terms, locating content of interest amounts to little more than guesswork. As the audiovisual content is only available to view for a short period of time, these subtitle files quickly become disconnected from their associated video files, thereby removing a means of accessing relevant contextual information. However, the realities of retroactively annotating records are, in many ways, no different from annotating contemporaneous records. The process of subject analysis remains the same for both historic and recent records, thereby making it possible to extract themes and translate these into controlled vocabulary terms.

When viewed in isolation, subtitle files can, however, be ambiguous in terms of speaker identification. Typically, the text itself will not explicitly say the name of the person speaking as this information is provided by the visual imagery. Although inferences can be made, this raises concerns about guesswork and to what extent metadata creators should input their own inferences into the metadata itself, risking the possibility of providing incorrect information. It seems, therefore, that we can retroactively annotate records from the past to the extent in which they are indexed with suitable subject terms. This provides more value than is currently provided by archived STV news subtitles. However, it is not always possible to assign speakers to text and, consequently, users will have to make their own inferences about this.

5.2.4. RQ4: What is the trade-off of implementing an accessible subtitle archive in a real-life context?

Almost every aspect of the annotation process comes with a trade-off in time investment. As discussed in the literature review, time plays a significant role when implementing metadata creation. For users, it can be time-consuming to search for content relevant to their interests. Digitisation offers potential for faster searching, but metadata needs to be created in order to make the content discoverable. The time investment needs to come from the people creating the metadata, in order to improve the user experience. However, as shown in this dissertation, annotating even a small sample of data can be a lengthy process.

I followed Balbi's (2011) all-in approach, which supports a broad policy of preservation and access. With regards to my data samples, I highlighted the complex interplay between the various items. If some of the items are not preserved, these connections could be removed. On the other hand, the time spent annotating every item may result in a trade-off in metadata quality. However, as discussed in 4.3.2.2, these could be treated as collections rather than discrete items. This makes more sense from a user perspective, and would save some time in creating metadata records.

There would also be an initial start-up cost in implementing an accessible subtitle archive. This would come from things like training, and policy development. It could also come from the time-investment associated with designing a small-scale controlled vocabulary. As discussed in section 4.2.3.1, this would be useful for specialised terminology. It would take time to create but, once established, would offer a more accurate way of describing certain content that is otherwise not well-served by existing controlled vocabularies. Thus, the trade-off is the initial time investment, but ultimately, this would provide more accurate resource description.

Miller (2011) notes that all metadata projects are subject to limitations. The people involved must perform to the best of their ability with the staff, budget and time they are able to invest in the project. In this sense, a project designed to improve accessibility to a subtitle archive is no different. If it were to be implemented in-house at STV, there be a small number of staff and very little spare time to spend on the project. STV is a business, and the primary focus of the subtitling department is providing an accessibility service to the deaf and hard of hearing. Consequently, the trade-off could be that the metadata is lacking in richness. Equally, the trade-off could come from the significant time taken before any end results are produced.

5.3. Recommendations

Based on the findings from the research questions, further investigation and experimentation into outsourcing some of the metadata creation is recommended. I discussed the idea of collaborative tagging in section 4.2.3.2. Due to the significant time investment needed to annotate both historic and recent data, it seems that there would need to be some form of outside help in order to make an accessible subtitle archive an achievable goal.

Alternatively, further research might consider other organisations who could take responsibility for maintaining the archive. As Snider (2000) argues, libraries already do this with traditional print news. Therefore, it could be a natural extension for them to take on the responsibility of a subtitle archive. I relied quite heavily on my personal knowledge of the organisation and news set-up throughout this study. It would be interesting to remove this ethnographic approach from the equation, and assess how somebody outside of the organisation would approach the creation of subtitle metadata.

These findings also highlighted certain items that were more complex to create metadata for. For example, interstitial material serves a different function from main news items. It would be useful for further exploration to be carried out on larger datasets to identify any further problematic cases. From this, policy guidelines could be developed in order to ensure these problematic cases are handled with clarity and consistency.

5.4. Conclusion

In conclusion, there is value to be gained from subtitle archive. Despite the potential for subtitling errors, broadly speaking my datasets seemed to provide an accurate account of the broadcast news. Apart from speaker identification, it was easy to read the subtitles and interpret the textual content, even when separated from the audiovisual footage. However, in order to make these subtitle files accessible, a significant time investment is required. In a professional-setting, other work commitments are likely to interfere with this, which may mean that it is some time before an accessible subtitle archive is possible.

This work is important because, to date, there are no studies which expanded on Snider's (2000) idea of creating a subtitle archive. This research took the first steps towards establishing this, by identifying areas of concerns and suggestions for improvement. Subtitles already serve an important function in terms of accessibility for deaf and hard of hearing audiences. It is hoped this research will encourage people to consider how subtitles' usage can be extended beyond their original purpose to provide additional value.

6.0. References

A Solution for Sharing News (2019) Available at: https://iptc.org/standards/nitf/ (Accessed: 10 June 2019).

A unique archive (2019) Available at: https://www.britishnewspaperarchive.co.uk/content/a_unique_archive (Accessed: 1 May 2019).

About ITV (2019) Available at: https://www.itvplc.com/about/history/2017 (Accessed: 28 July 2019).

About the archive (n.d.) Available at: https://tvnews.vanderbilt.edu/about (Accessed: 29 May 2019).

About Us (2019) Available at: https://www.redbeemedia.com/about-us/ (Accessed 6 August 2019).

Access Archives (2019) Available at: https://www.bbc.co.uk/informationandarchives/access_archives (Accessed 12 August 2019).

Action on Hearing Loss (2015) *Hearing Matters.* Available at: https://www.actiononhearingloss.org.uk/-/media/ahl/documents/research-and-policy/reports/hearing-matters-report.pdf (Accessed: 13 June 2019).

Action on Hearing Loss (2019) *Access to TV and Video on Demand (VOD) for people with hearing loss.* Available at: https://www.actiononhearingloss.org.uk/-/media/ahl/documents/research-and-policy/policy-statements/access-to-services-broadcasting-and-telecommunications/tv-and-vod-policy-statement-2018_final.pdf (Accessed: 11 August 2019).

Allen, R. and Johnson, K. (2008) ‘Preserving digital local news’. *The Electronic Library*, 26(3), pp. 387-398.

Allen, R. and Schalow, J. (1999) ‘Metadata and data structures for the historical newspaper digital library’, *CIKM 99 Proceedings of the eighth international conference on Information and knowledge management*. Kansas City, Missouri, 2 – 6 November 1999. Available at: http://boballen.info/RBA/PAPERS/CIKM1999/meta.pdf (Accessed: 4 March 2019).

Althaus, S., Edy, J. and Phalen, P. (2002) ‘Using the Vanderbilt Television Abstracts to Track Broadcast News Content: Possibilities and Pitfalls’. *Journal of Broadcasting & Electronic Media*, 46(3), pp. 473-492.

Amaral, R. and Trancoso, I. (2003) ‘Topic indexing of TV broadcast news programs’ in Mamede, N., Trancoso, I., Baptista, J. and das Graças Volpe Nunes, M. (eds.) *Computational Processing of the Portuguese Language*. Berlin: Springer. pp. 219-226.

American Library Association Committee on Cataloging (2000) ‘Task force on metadata’. Available at: https://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html (Accessed 12 June 2019).

Amernick, D. (2018) ‘The “Not Ready for Archive Players”: The Lost Seasons of Saturday Night Live’. *Journal of Popular Film and Television*, 46(2), pp. 70-81.

Andrade, M. and Baptista, A.A. (2015) ‘The use of application profiles and metadata schema by digital repositories: results from a survey’, *International Conference on Dublin Core and Metadata Applications.* Sao Paulo, Brazil, 1 – 4 September 2015. Available at: http://repositorio.ufes.br/jspui/bitstream/10/2022/1/362-1349-1-PB.pdf (Accessed: 8 August 2019).

Assmann, A. (2011) *Cultural memory and western civilization*. Cambridge: Cambridge University Press.

Assmann, I. and Mearns, M. (2015) ‘From broadcasting to archiving: the Southern African public service broadcast archives’. *Archives and Records*, 36(2), pp. 146-166.

Atton, C. (1998) ‘The librarian as ethnographer: notes towards a strategy for the exploitation of cultural collections’. *Collection Building*, 17(4), pp. 154-158.

Baez Montero, I.C. and Fernandez Soneira, A.M. (2010) ‘Spanish deaf people as recipients of closed captioning’, in Matamala, A. and Orero, P. (eds.) *Listening to subtitles*. Bern: Peter Lang, pp. 25-44.

Balbi, G. (2011) ‘Doing media history in 2050’. *Westminster Papers in Communication and Culture*, 8(1), pp. 115-133.

Beal-Alvarez, J. and Cannon, J.E. (2014) ‘Technology intervention research with deaf and hard of hearing learners: levels of evidence’. *American Annals of the Deaf,* 158(5), pp. 486-505.

Bingham, A. (2010) 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians'. *Twentieth Century British History*, 21(2), pp. 225-231.

Brown, M. *et al.* (1995) ‘Automatic content-based retrieval of broadcast news’, *ACM Multimedia 95*: *Proceedings of the third ACM international conference on Multimedia.* San Francisco, California, 2 – 6 November 1995. Available at: https://www.cl.cam.ac.uk/research/dtg/www/publications/public/files/paper.95.8 (Accessed: 4 March 2019).

Bryant, S. (2010) ‘National Television Archives and Their Role’. *Critical Studies in Television: The International Journal of Television Studies*, 5(2), pp. 60-67.

Butler, J. (2019) ‘Perspectives of deaf and hard of hearing viewers of captions’. *American Annals of the Deaf,* 163(5), pp. 534-553.

Can I use BBC content (2018) Available at: http://www.bbc.co.uk/terms/can-i-use-bbc-content (Accessed: 13 August 2019).

Cigognetti, L. (2001) ‘Historians and TV archives’, in Roberts, G. and Taylor, P.M. (eds.) *The Historian, Television and Television History.* Luton: University of Luton Press, pp. 33-38.

Clair, K. (2008) ‘Developing an audiovisual metadata application profile’. *Library Collections, Acquisitions, & Technical Services*, 32(1), pp. 53-57.

Cohen, D.J. and Rosenzweig, R. (2006) *Digital history: a guide to gathering, preserving, and presenting the past on the web.* Philadelphia, Pennsylvania: University of Philadelphia Press.

Collins, K. (2010) ‘The trouble with Archie: locating and accessing primary sources for the study of the 1970s US sitcom, All in the Family’. *Critical Studies in Television: The International Journal of Television Studies*, 5(2), pp. 118-132.

Compton, M.A. (2007) ‘The archivist, the scholar, and access to historic television materials’. *Cinema Journal,* 46(3), pp. 129-133.

Council of Europe (2001) *European convention for the protection of the audiovisual heritage.* Available at: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168008155f (Accessed: 1 May 2019).

Cox, M., Tadic, L. and Mulder, E. (2006) *Descriptive metadata for television: an end-to-end introduction.* Burlington, USA: Focal Press.

Darlington, J., Finney, A. and Pearce, A. (2003) ‘Domesday Redux: The rescue of the BBC Domesday Project videodiscs’. *Microform & Imaging Review*, 32(4), pp. 113-118.

Demiros, I. *et al*. (2008) ‘Media Monitoring by Means of Speech and Language Indexing for Political Analysis’. *Journal of Information Technology & Politics*, 5(1), pp. 133-146.

De Sutter, Notebaert and Van de Walle (2006) ‘Evaluation of metadata standards in the context of audio-visual libraries’, in *Research and advanced technologies for digital libraries.* Alicante, Spain, 17-22 September 2006. Available at: https://s3.amazonaws.com/academia.edu.documents/31081454/34.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1558606687&Signature=ile3cGYlSfkt7oHpr2YD29uXAWw%3D&response-content-disposition=inline%3B%20filename%3DThe_SINAMED_and_ISIS_projects_applying_t.pdf#page=236 (Accessed: 23 May 2019).

Documentation for NITF (n.d.) Available at: https://www.iptc.org/std/NITF/3.5/documentation/nitf.html (Accessed: 1 August 2019).

Dowman, M. *et al*. (2005) ‘Web-assisted annotation, semantic indexing and search of television and radio news’, *WWW ’05 Proceedings of the 14th international conference on World Wide Web*. Chiba, Japan, 10-14 May 2005. Available at: https://dl.acm.org/citation.cfm?doid=1060745.1060781 (Accessed 2 May 2019).

Downey, G.J. (2008) *Closed captioning: subtitling, stenography, and the digital convergence of text with television.* Baltimore, Maryland: The John Hopkins University Press.

Dragon Speech Recognition Software (2019) Available at: https://www.nuance.com/en-gb/dragon.html (Accessed: 20 June 2019).

Erickson, C. (2013) *Light, dim and dark archives: what are they?* Available at: http://preservationmatters.blogspot.com/2013/05/light-dark-and-dim-archives-what-are.html (Accessed: 1 August 2019).

European Commission (2008) *Copyright in the knowledge economy.* Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0466:FIN:EN:PDF (Accessed: 1 May 2019).

Feather, J. (2018) 'Introduction: principles and policies', in Feather, J. (ed.) *Managing preservation for libraries and archives*. London: Routledge.

Fenton, C. (2010) 'Use of Controlled Vocabulary and Thesauri in UK Online Finding Aids'. *Journal of the Society of Archivists*, 31(2), pp. 187-205.

Fleming, P. and King, E. (2009) 'The British Library newspaper collections and future strategy'. *Interlending & Document Supply*, 37(4), pp. 223-228.

Gaber, I. (1998) 'Television and political coverage', in Geraghty, C. and Lusted, D. (eds.) *The Television Studies Book*. London: Arnold, pp. 264-273.

Glossary of Broadcast Terms (2011) Available at: https://centreforjournalism.co.uk/content/tv-studio-terms (Accessed: 11 June 2019).

Guy, M. and Tonkin, E. (2006) 'Folksonomies: Tidying up tags?'. *D-Lib Magazine*, 12(1).

Hale, M., Fowler, E. and Goldstein, K. (2007) 'Capturing Multiple Markets: A New Method of Capturing and Analyzing Local Television News'. *Electronic News*, 1(4), pp. 227-243.

Hansen, K. and Paul, N. (2015) 'Newspaper archives reveal major gaps in digital age'. *Newspaper Research Journal*, 36(3), pp. 290-298.

Haynes, D. (2018) *Metadata for information management and retrieval.* 2nd edn. London: Facet.

Heeren, W., Ordelman, R. and de Jong, F. (2008) 'Affordable access to multimedia by exploiting collateral data', *Proceedings of CBMI 2008.* London, UK, 18-20 June 2008. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4564994 (Accessed: 29 April 2019).

Hewett, R. (2014) *Academic requirements for pre-1989 BBC archive content.* Available at: https://www.researchgate.net/profile/Richard_Hewett2/publication/301499469_Academic_Requirements_for_Pre-1989_BBC_Archive_Content/links/5717e86108aed8a339e5b11a/Academic-Requirements-for-Pre-1989-BBC-Archive-Content.pdf (Accessed: 10 June 2019).

Hillmann, D. (2005) *Using Dublin Core*. Available at: http://dublincore.org/documents/usageguide/ (Accessed 1 July 2019).

Holley, R. (2010) ‘Tagging full text searchable articles: an overview of social tagging activity in historic Australian newspapers August 2008-August 2009’. *D-Lib Magazine*, 16(1/2).

Howarth, L. (2003) ‘Designing a Common Namespace for Searching Metadata-Enabled Knowledge Repositories: An International Perspective’. *Cataloging & Classification Quarterly*, 37(1-2), pp. 173-185.

Introduction to the BFI Collections (2019) Available at: https://www.bfi.org.uk/archive-collections/introduction-bfi-collections/exploring-collections/television (Accessed: 19 June 2019).

ITV Archive (n.d.) Available at: https://www.itvarchive.com/about-ext/ (Accessed: 12 August 2019).

Jensema, C. (1997) *Viewer reaction to different captioned television speeds.* PhD thesis. Institute for Disabilities Research and Training, Inc. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.575.4056&rep=rep1&type=pdf (Accessed: 30 May 2019).

Joudrey, D., Taylor, A. and Miller, D. (2015) *Introduction to cataloging and classification*. Santa Barbara, California: Libraries Unlimited.

Kang, S., Gearhart, S. and Bae, H. (2010) ‘Coverage of Alzheimer’s Disease From 1984 to 2008 in Television News and Information Talk Shows in the United States: An Analysis of News Framing’. *American Journal of Alzheimer's Disease & Other Dementias*, 25(8), pp. 687-697.

Kaufman, P.A. *et al*. (1993) ‘Why going online for content analysis can reduce research reliability”. *Journalism Quarterly,* 70(4), pp. 824-832.

Kernell, G., Lamberson, P. and Zaller, J. (2017) ‘Market Demand for Civic Affairs News’. *Political Communication*, 35(2), pp. 239-260.

Khoo, M. and Hall, C. (2013) ‘Managing metadata: Networks of practice, technological frames, and metadata work in a digital library’. *Information and Organization*, 23(2), pp. 81-106.

King, E. (2005) ‘Digitisation of Newspapers at the British Library’. *The Serials Librarian*, 49(1-2), pp. 165-181.

Kramp, L. (2014) ‘Media studies without memory? Institutional, economic and legal issues of accessing television heritage in the digital age’, in Kramp, L. et al. (eds.) *Media Practice and Everyday Agency in Europe*. Bremen: edition lumiere, pp. 227-248. Available at: https://www.researchgate.net/profile/Nico_Carpentier/publication/268980708_Media_Practice_and_Everyday_Agency_in_Europe/links/547c95e50cf2cfe203c1f21e.pdf#page=226 (Accessed 29 April 2019).

Krippendorff, K. (1980) *Content Analysis.* California: Sage.

Kuklinski, J. and Sigelman, L. (1992) ‘When Objectivity is Not Objective: Network Television News Coverage of U.S. Senators and the "Paradox of Objectivity."’ *The Journal of Politics*, 54(3), pp. 810-833.

Kyle, J.G. (1993) *Switched on… or not? Deaf people’s views on television subtitling.* Bristol: The Deaf Studies Trust. Available at: http://www.deafstudiestrust.org/files/pdf/reports/switched_on_summary1993.pdf (Accessed: 13 June 2019).

Lessig, L. (2004) *Free Culture.* Petter Reinholdtsen: Oslo.

Library of Congress (2016) ‘How do we determine aboutness?’. Available at: https://www.loc.gov/catworkshop/lcsh/PDF%20scripts/1-4-Aboutness.pdf (Accessed: 20 June 2019).

Looking for old programmes (2017) Available at: https://www.transdiffusion.org/about/looking-for-old-programmes/ (Accessed: 1 July 2019).

Ma, J. (2006) ‘Managing metadata for digital projects’. *Library Collections, Acquisitions, & Technical Services*, 30(1-2), pp. 3-17.

Macgregor, G. and McCulloch, E. (2006) ‘Collaborative tagging as a knowledge organisation and resource discovery tool’. *Library Review*, 55(5), pp. 291-300.

Madden, P. (1981) *Keeping Television Alive: The Television Work of the National Film Archive.* London: BFI.

Maguire, B. (2002) ‘Television network news coverage of corporate crime from 1970-2000’. *Western Criminology Review*, 3(2), pp. 1-22.

Marshall, C. (1998) ‘Making metadata: a study of metadata creation for a mixed physical-digital collection’, *Proceedings of the third ACM conference on digital libraries.* Pittsburgh, Pennsylvania, 23-26 June 1998. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.442.9039&rep=rep1&type=pdf (Accessed: 23 May 2019).

Martin, J. (2005) ‘The Dawn of Tape: Transmission Device as Preservation Medium’. *The Moving Image*, 5(1), pp. 45-66.

Massis, B. (2012) ‘Local newspapers and the library: a “community asset”’. *New Library World*, 113(11/12), pp. 614-618.

Maurantonio, N. (2014) ‘Archiving the Visual’. *Media History*, 20(1), pp. 88-102.

McClure, R.R. (1999) *Abstracts and transcriptions.* Available at: https://www.genealogy.com/articles/twigs/rhonda072999.html (Accessed: 13 August 2019).

McCutcheon, S. (2009) ‘Keyword vs controlled vocabulary searching: the one with the most tools wins’. *The Indexer: The International Journal of Indexing*, 27(2), pp. 62-65.

McIvor, J. (2012) *Broadcaster STV reaches new deal with ITV*. Available at: https://www.bbc.co.uk/news/uk-scotland-scotland-business-17258458 (Accessed: 28 July 2019).

McKernan, L. (2014) *10 great online newspaper archives.* Available at: https://blogs.bl.uk/thenewsroom/2014/02/10-great-online-newspaper-archives.html (Accessed: 15 May 2019).

MediaCentral (2019) Available at: https://www.avid.com/products/mediacentral (Accessed: 25 June 2019).

Media Topics (2019) Available at: https://iptc.org/standards/media-topics/ (Accessed: 30 June 2019).

Messina, A. *et al*. (2006) 'Creating Rich Metadata in the TV Broadcast Archives Environment: The PrestoSpace Project', *2006 Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS'06)*. Leeds, UK, 13-15 December 2006. Available at: https://www.researchgate.net/profile/Walter_Allasia/publication/221251586_Creating_Rich_Metadata_in_the_TV_Broadcast_Archives_Environment_The_PrestoSpace_Project/links/57726de108aeef01a0b62b3d/Creating-Rich-Metadata-in-the-TV-Broadcast-Archives-Environment-The-PrestoSpace-Project.pdf (Accessed: 12 May 2019).

Miller, S. J. (2011) *Metadata for digital collections*. London: Facet.

Murphy, W. (1997) *Television and video preservation 1997. A report on the current state of American television and video preservation.* Washington, DC: Library of Congress. Available at: http://www.loc.gov/static/programs/national-film-preservation-board/documents/tvstudy.pdf (Accessed: 1 May 2019).

Mussell, J. (2017) 'Beyond the great index: digital resources and actual copies', in Shattock, J. (ed.) *Journalism and the Periodical Press in Nineteenth-Century Britain.* Cambridge: Cambridge University Press, pp. 17-30.

National Information Standards Organization. (2010) *Guidelines for the Construction, Format and Management of Monolingual Controlled Vocabularies*. Baltimore: National Information Standards Organization. Available at: https://groups.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf (Accessed: 10 July 2019).

Neal, D. (2008) 'News Photographers, Librarians, Tags, and Controlled Vocabularies: Balancing the Forces'. *Journal of Library Metadata*, 8(3), pp. 199-219.

NewsML-G2 (2019) Available at: https://iptc.org/standards/newsml-g2/ (Accessed: 1 August 2019).

Neves, J. (2005) *Audiovisual translation: subtitling for the deaf and hard-of-hearing.* PhD thesis. Roehampton University. Available at:

https://iconline.ipleiria.pt/bitstream/10400.8/409/1/Thesis%20agosto%202005.pdf (Accessed: 10 June 2019).

Nicholson, B. (2013).'The digital turn'. *Media History*, 19(1), pp. 59-73.

Norris, P. (1995) 'The restless searchlight: Network news framing of the post-Cold War world'. *Political Communication*, 12(4), pp. 357-370.

O'Connor, D. (2015) 'Extending and enriching the official publications collection at the UCL institute of education: Developing, maintaining and enhancing the digital education resource archive'. *Refer, 31*(2), pp. 29-31.

Ofcom (2006) 'Television access services'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0016/42442/access.pdf (Accessed: 13 June 2019).

Ofcom (2017a) 'Ofcom's code on television access services'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0020/97040/Access-service-code-Jan-2017.pdf (Accessed: 13 June 2019).

Ofcom (2017b) 'The communications market: Scotland'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0020/104933/cmr-2017-scotland.pdf (Accessed: 15 June 2019).

Ofcom (2018a) 'Media nations: Scotland 2018'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0011/116012/media-nations-2018-scotland.pdf (Accessed: 12 May 2019).

Ofcom (2018b) 'Communications market report'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf (Accessed: 12 May 2019).

Ofcom (2018c) 'News consumption in the UK: 2018'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0024/116529/news-consumption-2018.pdf (Accessed: 16 May 2019).

Ofcom (2018d) 'The changing worlds of news: qualitative research'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0023/115916/The-Changing-World-of-News.pdf (Accessed: 20 June 2019).

Olson, H. (2001) 'The Power to Name: Representation in Library Catalogs'. *Signs: Journal of Women in Culture and Society*, 26(3), pp. 639-668.

Palinkas, L. *et al*. (2013) 'Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research'. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5), pp. 533-544.

Park, J. and Childress, E. (2009) 'Dublin Core metadata semantics: an analysis of the perspectives of information professionals'. *Journal of Information Science*, 35(6), pp. 727-739.

Patton, M. (2002) *Qualitative Research & Evaluation Methods.* London: Sage.

PBCore vocabularies (n.d.) Available at: https://pbcore.org/pbcore-controlled-vocabularies/creatorrole-and-contributorrole-vocabulary/ (Accessed: 19 July 2019).

Pickard, A. (2008) *Research methods in information*. London: Facet.

Pribble, J.M. *et al*. (2006) 'Medical news for the public to use: what's on local TV news'. *American Journal of Managed Care,* 12(3), pp. 170-176.

Ragaller, I. and Rafferty, P. (2012) 'Biases in the classification of Welsh art material'. *Aslib Proceedings*, 64(3), pp. 262-273.

Ramesh, P., Vivekavardhan, J. and Bharathi, K. (2015) 'Metadata Diversity, Interoperability and Resource Discovery Issues and Challenges'. *DESIDOC Journal of Library & Information Technology*, 35(3), pp. 193-199.

Ranjan, A., Balakrishnan, R. and Chignell, M. (2006) 'Searching in audio: the utility of transcripts, dichotic presentation and time-compression', *Proceedings of CHI 2006.* Quebec, Canada, 22-27 April 2006. Available at: https://www.researchgate.net/profile/Mark_Chignell/publication/221515644_Searching_in_audio_The_utility_of_transcripts_dichotic_presentation_and_time-compression/links/5714c88208aeebe07c06c4fb/Searching-in-audio-The-utility-of-transcripts-dichotic-presentation-and-time-compression.pdf (Accessed: 29 April 2019).

Razikin, K. *et al*. (2011) 'Social tags for resource discovery: a comparison between machine learning and user-centric approaches'. *Journal of Information Science*, 37(4), pp. 391-404. Available at: https://journals.sagepub.com/doi/pdf/10.1177/0165551511408847 (Accessed: 27 May 2019).

Ribas, M.A. and Romero Fresco, P. (2008) 'A practical proposal for the training of respeakers'. *The Journal of Specialised Translation*, 10, pp. 106-127.

Riley, J. (2017) *Understanding metadata: what is metadata, and what is it for? A primer.* Baltimore, MD: National Information Standards Organization (NISO). Available at: https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf (Accessed: 29 June 2019).

Robson, G.D. (2004) *The Closed Captioning Handbook.* Amsterdam: Focal Press.

Romer, D., Jamieson, P. and Jamieson, K. (2006) 'Are News Reports of Suicide Contagious? A Stringent Test in Six U.S. Cities'. *Journal of Communication*, 56(2), pp. 253-270.

Rooks, S. (2010) 'What Happened to the BBC Sound Archive?' *Journal of the Society of Archivists*, 31(2), pp. 177-185.

Ryan, D. and Simon, J. (2014) 'Broadcast news transcripts in academic news databases'. *Center for Research Libraries*, 34(1).

Sample records (n.d.) Available at: https://pbcore.org/sample-records/description-document-with-instantiations (Accessed: 1 August 2019).

Schuller, D. (2015) 'Socio-Technical and Socio-Cultural Challenges of Audio and Video Preservation'. *International Preservation News,* (46), pp. 5-7.

Screen Systems (2015) Available at: https://subtitling.com/downloads/wincaps-q-live-brochure/?wpdmdl=6342 (Accessed: 10 June 2019).

Snider, J.H. and Janda, K. (1998) 'Newspapers in bytes and bits: limitations of electronic databases for content analysis', *Annual Meeting of the American Political Science Association*. Boston, 3-6 September 1998. Available at: https://www.researchgate.net/profile/Kenneth_Janda/publication/237103675_Newspapers_in_Bytes_and_Bits_Limitations_of_Electronic_Databases_for_Content_Analysis/links/0deec53c5430295f1b000000/Newspapers-in-Bytes-and-Bits-Limitations-of-Electronic-Databases-for-Content-Analysis.pdf (Accessed: 1 June 2019).

Snider, J.H. (2000) 'Local TV news archives as a public good'. *The Harvard International Journal of Press/Politics,* 5(2), pp. 111-117.

Spigel, L. (2005) 'Our TV heritage: television, the archive, and the reasons for preservation', in Wasko, J. (ed.) *A Companion to Television.* Oxford, UK: Blackwell, pp. 67-99.

Strottman, T. (2007) 'Some of Our Fifty Are Missing: Library of Congress Subject Headings for Southwestern Cultures and History'. *Cataloging & Classification Quarterly*, 45(2), pp. 41-64.

STV (2011) *Annual report and accounts 2011.* Available at: http://www.stvplc.tv/files/download/4ec1e887b214b33 (Accessed: 10 June 2019).

STV (2017) *Annual report and accounts 2017.* Available at: http://www.stvplc.tv/files/download/d8e337836f012fa (Accessed: 10 June 2019).

STV (2018) *Annual report and accounts 2018.* Available at: http://www.stvplc.tv/files/download/2f9039564bc16b2 (Accessed: 12 May 2019).

Suominen, O. and Mader, C. (2013) 'Assessing and Improving the Quality of SKOS Vocabularies'. *Journal on Data Semantics*, 3(1), pp. 47-73.

Tanackovic, S.F., Krtalic, M. and Lacovic, D. (2014) 'Newspapers as a research source: information needs and information seeking of humanities scholars', *IFLA 2014*. Lyon, France, 16-22 August 2014. Available at: https://cf-www.ifla.org/files/assets/newspapers/Geneva_2014/s6-lacovic-en.pdf (Accessed: 2 May 2019).

Ubois, J. (2006) 'Finding Murphy Brown: How accessible are historic television broadcasts?' *Journal of Digital Information*, 7(2), pp. 1-20.

UKAT (2019) Available at: https://ukat.aim25.com/ (Accessed: 15 July 2019).

Wactlar, H.D. and Christel, M.G. (2002) 'Digital video archives: managing through metadata', in in Marcum, D. and Campbell, L. *Building a national strategy for digital preservation: Issues in digital media archiving*. Washington: Council on Library and Information Resources and Library of Congress, pp. 80-93.

Walsh, J. (2011) 'The use of Library of Congress Subject Headings in digital collections'. *Library Review*, 60(4), pp. 328-343.

Washington, A. and Weidner, A. (2017) 'Collaborative metadata application profile development for DAMS migration', *International Conference on Dublin Core and Metadata Applications.* Washington, DC, USA, 26 – 19 October 2017. Available at: https://uh-ir.tdl.org/handle/10657/2069 (Accessed: 8 August 2019).

We Are TVT Media (n.d.) Available at: https://www.tvt.media/who-we-are/ (Accessed: 6 August 2019).

Weibel, S. and Miller, E. (2001) 'Image Description on the Internet'. *Journal of Library Administration*, 34(1-2), pp. 209-219.

Whannel, G. (2005) 'Pregnant with anticipation'. *International Journal of Cultural Studies*, 8(4), pp. 405-426.

What are tags for (2019) Available at: https://www.britishnewspaperarchive.co.uk/help-faq/what-are-tags-for (Accessed: 4 May 2019).

What is PBCore (n.d.) Available at: https://pbcore.org/ (Accessed: 19 July 2019).

Wisneski, R. and Dressler, V. (2009) 'Implementing TEI Projects and Accompanying Metadata for Small Libraries: Rationale and Best Practices'. *Journal of Library Metadata*, 9(3-4), pp. 264-288.

Wright, R. (2009) 'Preservation of broadcast archives – a BBC perspective.' *International Presrvation News*, 47, pp. 13-17.

XMLNews Technical Overview (1999) Available at: http://www.xmlnews.org/docs/tech-overview.html (Accessed: 10 June 2019).

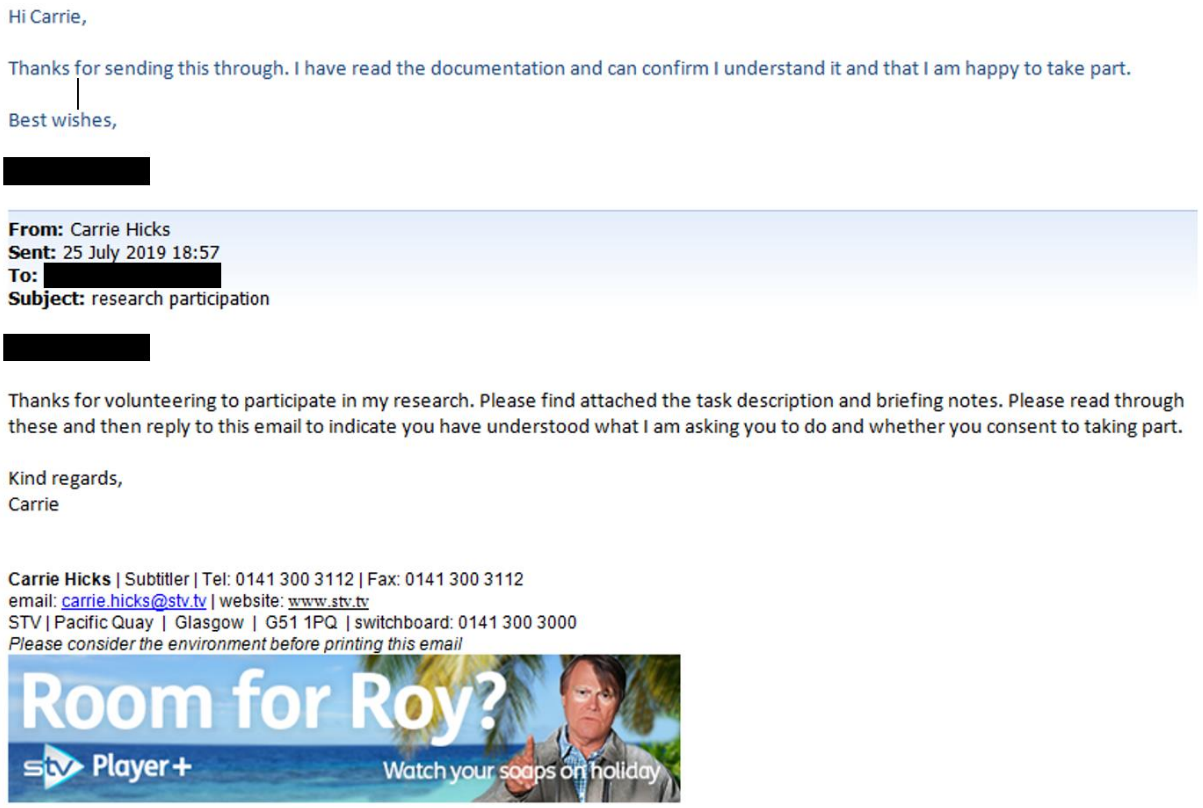
XMLNews-Meta Technical Specification (1999) Available at: http://www.xmlnews.org/docs/meta-spec.html (Accessed: 20 July 2019).

XML Tree (2019) Available at: https://www.w3schools.com/xml/xml_tree.asp (Accessed: 26 July 2019).

Yaginuma, T., Mendes Pereira, T.S. and Baptista, A.A. (2003) 'Metadata elements for digital news resource description', *Congresso Luso-Mocambicano de Engenharia.* Maputo, Mozambique, 19-21 August 2003. Available at: https://repositorium.sdum.uminho.pt/bitstream/1822/279/1/CLME2003_1317-1326.pdf (Accessed: 12 June 2019).

Zeng, M. and Qin, J. (2008) *Metadata*. New York: Neal-Schuman Publishers.

Appendix A: Ethics approval



Hi Carrie,

Thanks for sending this through. I have read the documentation and can confirm I understand it and that I am happy to take part.

Best wishes,

[redacted]

From: Carrie Hicks
Sent: 25 July 2019 18:57
To: [redacted]
Subject: research participation

[redacted]

Thanks for volunteering to participate in my research. Please find attached the task description and briefing notes. Please read through these and then reply to this email to indicate you have understood what I am asking you to do and whether you consent to taking part.

Kind regards,
Carrie

Carrie Hicks | Subtitler | Tel: 0141 300 3112 | Fax: 0141 300 3112
email: carrie.hicks@stv.tv | website: www.stv.tv
STV | Pacific Quay | Glasgow | G51 1PQ | switchboard: 0141 300 3000
Please consider the environment before printing this email

Task description:

Please look at the following news items shown below, and then select appropriate subject categories from the Media Topics controlled vocabulary, which you can access here: http://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

Please choose at least 2 subject categories for each item.

Title	G-JOHNSTON
Subject	

Title	G1-TENNIS
Subject	

Title	G1-POLL
Subject	

Title	G2-POLL
Subject	

Title	G1-BIOMETRIC
Subject	

Title	G2-BIOMETRIC
Subject	

Briefing notes

Title of research project: Organisational impact of an archive of news subtitles: usefulness and accessibility

Supervisor: Dr Martin Halvey

Purpose of the research: The purpose of this research is to find out how to improve the accessibility of a news subtitle archive, and to explore what value this would have. The aim is to enhance discoverability of individual news items, including historic data samples.

Data collection and handling: As part of my dissertation, I have applied subject categories to individual news stories in order to enhance their discoverability. To do this, I used an existing controlled vocabulary called IPTC Media Topics. I am using intercoder reliability to compare my results with someone else's. You will be required to look at a sample of news stories and assign subject categories to them using the IPTC Media Topics controlled vocabulary. This activity is not expected to take more than 30 minutes, and will be arranged at a time that suits you.

Confidentiality and anonymity: You are guaranteed total confidentiality with regard to anything you do, say or write in relation to this research. You will not be asked to look at any data that will distress you in any way. All data will be identified by a fictional name that is only known to the researcher. All data will be stored at my home and destroyed at the end of the formal period of retention. The only people with access to this data are my supervisor and examiners, who may request to see it but in no way will they attempt to identify you through this data.

Voluntary involvement: You are free to end your participation in this research at any time, and you may refuse to answer any questions or take part in any activity you do not wish to engage in. You also have the right to request I destroy or remove the data at any point from my study before it is submitted.

Contact details:

Secretary to the Departmental Ethics Committee

Department of Computer and Information Sciences

Livingstone Tower

Richmond Street

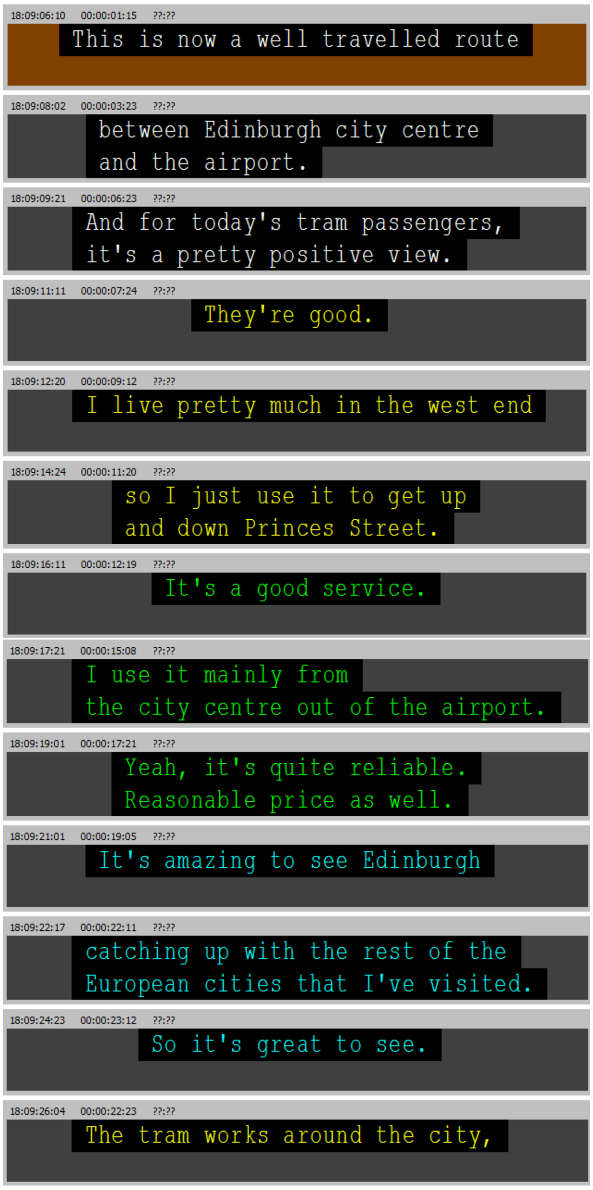
Glasgow

G1 1XH

Email: ethics@cis.strath.ac.uk

Appendix B: Subtitle examples

The images below show an example of a single news package. Due to space limitations, they are displayed as four separate images. Please read in the order: figure 32, figure 33, figure 34, figure 35.



18:09:06:10 00:00:01:15 ??:??
This is now a well travelled route

18:09:08:02 00:00:03:23 ??:??
between Edinburgh city centre
and the airport.

18:09:09:21 00:00:06:23 ??:??
And for today's tram passengers,
it's a pretty positive view.

18:09:11:11 00:00:07:24 ??:??
They're good.

18:09:12:20 00:00:09:12 ??:??
I live pretty much in the west end

18:09:14:24 00:00:11:20 ??:??
so I just use it to get up
and down Princes Street.

18:09:16:11 00:00:12:19 ??:??
It's a good service.

18:09:17:21 00:00:15:08 ??:??
I use it mainly from
the city centre out of the airport.

18:09:19:01 00:00:17:21 ??:??
Yeah, it's quite reliable.
Reasonable price as well.

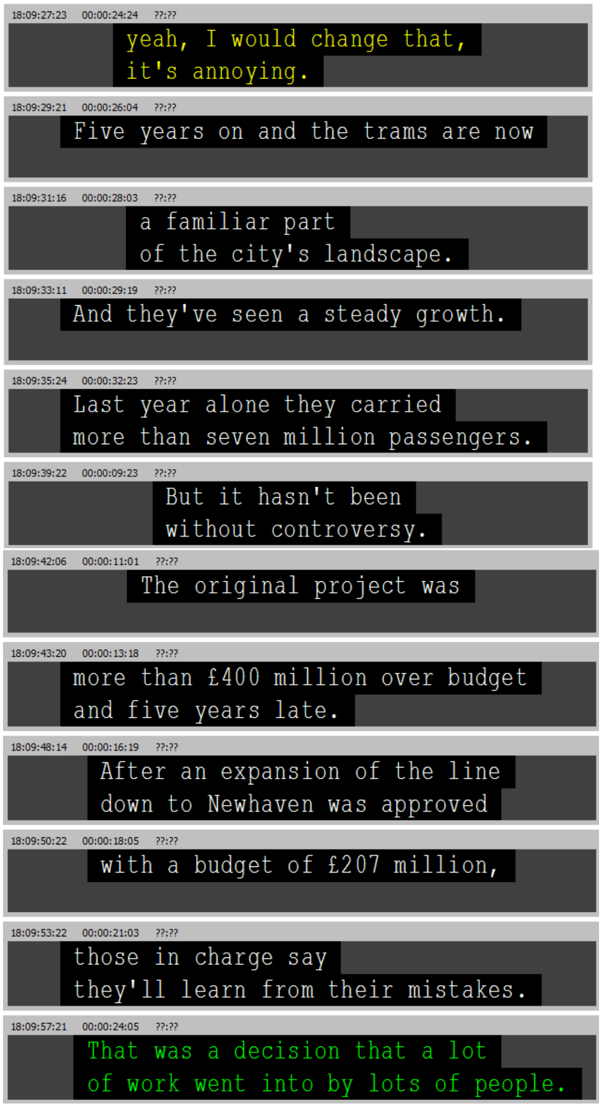
18:09:21:01 00:00:19:05 ??:??
It's amazing to see Edinburgh

18:09:22:17 00:00:22:11 ??:??
catching up with the rest of the
European cities that I've visited.

18:09:24:23 00:00:23:12 ??:??
So it's great to see.

18:09:26:04 00:00:22:23 ??:??
The tram works around the city,

Figure 33: Part 1



18:09:27:23 00:00:24:24 ??:??
yeah, I would change that,
it's annoying.

18:09:29:21 00:00:26:04 ??:??
Five years on and the trams are now

18:09:31:16 00:00:28:03 ??:??
a familiar part
of the city's landscape.

18:09:33:11 00:00:29:19 ??:??
And they've seen a steady growth.

18:09:35:24 00:00:32:23 ??:??
Last year alone they carried
more than seven million passengers.

18:09:39:22 00:00:09:23 ??:??
But it hasn't been
without controversy.

18:09:42:06 00:00:11:01 ??:??
The original project was

18:09:43:20 00:00:13:18 ??:??
more than £400 million over budget
and five years late.

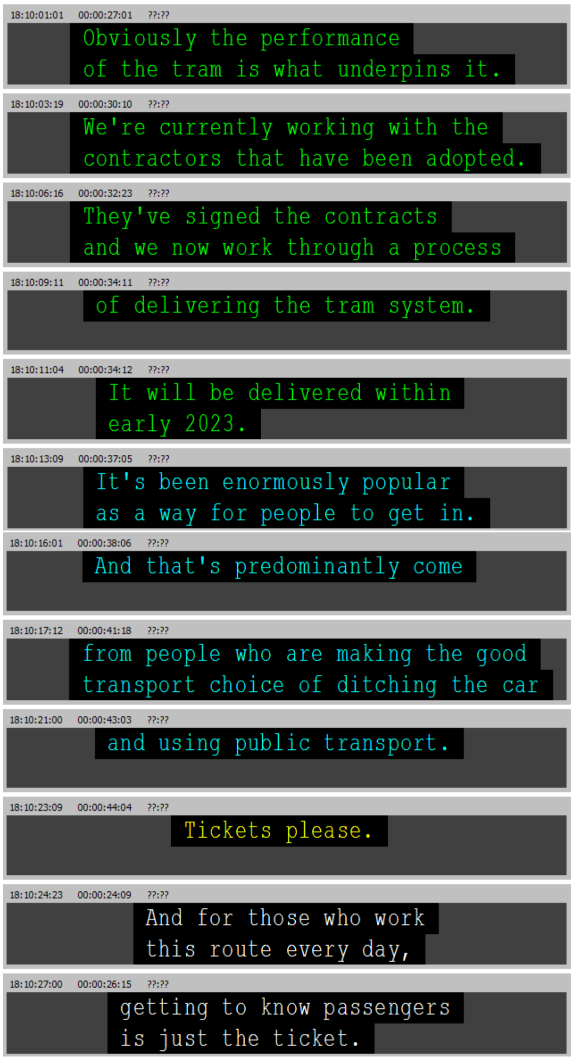
18:09:48:14 00:00:16:19 ??:??
After an expansion of the line
down to Newhaven was approved

18:09:50:22 00:00:18:05 ??:??
with a budget of £207 million,

18:09:53:22 00:00:21:03 ??:??
those in charge say
they'll learn from their mistakes.

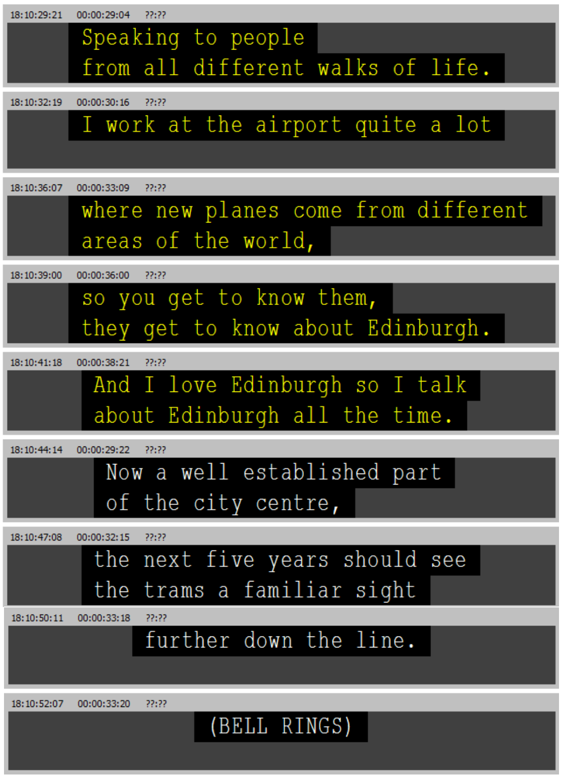
18:09:57:21 00:00:24:05 ??:??
That was a decision that a lot
of work went into by lots of people.

Figure 32: Part 2



```
18:10:01:01   00:00:27:01   ??:??
Obviously the performance
of the tram is what underpins it.

18:10:03:19   00:00:30:10   ??:??
We're currently working with the
contractors that have been adopted.

18:10:06:16   00:00:32:23   ??:??
They've signed the contracts
and we now work through a process

18:10:09:11   00:00:34:11   ??:??
of delivering the tram system.

18:10:11:04   00:00:34:12   ??:??
It will be delivered within
early 2023.

18:10:13:09   00:00:37:05   ??:??
It's been enormously popular
as a way for people to get in.

18:10:16:01   00:00:38:06   ??:??
And that's predominantly come

18:10:17:12   00:00:41:18   ??:??
from people who are making the good
transport choice of ditching the car

18:10:21:00   00:00:43:03   ??:??
and using public transport.

18:10:23:09   00:00:44:04   ??:??
Tickets please.

18:10:24:23   00:00:24:09   ??:??
And for those who work
this route every day,

18:10:27:00   00:00:26:15   ??:??
getting to know passengers
is just the ticket.
```

Figure 34: Part 3



```
18:10:29:21   00:00:29:04   ??:??
Speaking to people
from all different walks of life.

18:10:32:19   00:00:30:16   ??:??
I work at the airport quite a lot

18:10:36:07   00:00:33:09   ??:??
where new planes come from different
areas of the world,

18:10:39:00   00:00:36:00   ??:??
so you get to know them,
they get to know about Edinburgh.

18:10:41:18   00:00:38:21   ??:??
And I love Edinburgh so I talk
about Edinburgh all the time.

18:10:44:14   00:00:29:22   ??:??
Now a well established part
of the city centre,

18:10:47:08   00:00:32:15   ??:??
the next five years should see
the trams a familiar sight

18:10:50:11   00:00:33:18   ??:??
further down the line.

18:10:52:07   00:00:33:20   ??:??
(BELL RINGS)
```

Figure 35: Part 4

Appendix C: Metadata records

Metadata records for 2019 dataset (Items 1-32)

1

TITLE	G-OPENERS
IDENTIFIER	NEWSATSIXCEN310519G-OPENERS
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Intro to 6pm central news programme
SUBJECT	Sex crime; police; soccer; library and museum
SUBJECT	Celtic Football Club; forensics; Neil Lennon; Dumfries; national centre for reading
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
CONTRIBUTOR	Joanne Lumley
CONTRIBUTOR ROLE	Speaker

2

TITLE	G1-ABUSE
IDENTIFIER	NEWSATSIXCEN310519 G1-ABUSE
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Celtic are criticised for not taking responsibility over Celtic Boys Club coaches convicted of child sex abuse
SUBJECT	Sex crime; soccer
SUBJECT	Celtic Football Club; Celtic; Celtic FC; Celtic Boys Club; Peter Lawwell; Scottish Football Association
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G2-ABUSE

3

TITLE	G2-ABUSE
IDENTIFIER	NEWSATSIXCEN310519 G2-ABUSE
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Report on Celtic Boys Club coaches convicted of child sex abuse
SUBJECT	Sex crime; soccer
SUBJECT	Celtic Football Club; Celtic; Celtic FC; Celtic Boys Club; Peter Lawwell; Scottish Football Association; James Dornan; Adam Tomkins
CREATOR	Louise Scott
CREATOR ROLE	Reporter
CONTRIBUTOR	James Dornan
CONTRIBUTOR ROLE	Speaker
CONTRIBUTOR	Adam Tomkins
CONTRIBUTOR ROLE	Speaker
CONTRIBUTOR	Peter Lawwell
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G1-ABUSE

4

TITLE	G1-BIOMETRIC
IDENTIFIER	NEWSATSIXCEN310519G1-BIOMETRIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Creation of a proposed new law to oversee how personal forensic evidence is handled by police
SUBJECT	Science and technology; police; personal data collection; identification technology
SUBJECT	Forensics; forensic evidence; forensic technology
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G2-BIOMETRIC

5

TITLE	G2-BIOMETRIC
IDENTIFIER	NEWSATSIXCEN310519G2-BIOMETRIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Advances in forensic technology have raised concerns about how police handle the information so a new Scottish biometrics commissioner is proposed to oversee this.
SUBJECT	Science and technology; police; personal data collection; identification technology
SUBJECT	Forensics; forensic evidence; forensic technology; biometrics; Scottish biometrics commissioner
CREATOR	Ewan Petrie
CREATOR ROLE	Reporter
CONTRIBUTOR	Humza Yousaf

CONTRIBUTOR ROLE	Speaker
ISPARTOF	G1-BIOMETRIC

6

TITLE	G1-TRAM
IDENTIFIER	NEWSATSIXCEN310519G1-TRAM
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Proposed extension to the Edinburgh tram line
SUBJECT	Road transport; commuting
SUBJECT	Trams; Edinburgh; Newhaven
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G1A-TRAM
HASPART	G2-TRAM

7

TITLE	G1A-TRAM
IDENTIFIER	NEWSATSIXCEN310519G1A-TRAM
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Vanessa Kennedy reports live from on board a tram, talking about the planned tram extension.
SUBJECT	Road transport; commuting
SUBJECT	Trams; Edinburgh; Newhaven
CREATOR	Vanessa Kennedy
CREATOR ROLE	Reporter
HASPART	G2-TRAM
ISPARTOF	G1-TRAM

8

TITLE	G2-TRAM
IDENTIFIER	NEWSATSIXCEN310519G2-TRAM
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Today marks five years since the introduction of the over budget Edinburgh trams project
SUBJECT	Road transport; commuting; budgets and budgeting
SUBJECT	Trams; Edinburgh; Newhaven; over budget
CREATOR	Vanessa Kennedy
CREATOR ROLE	Reporter
CONTRIBUTOR	Adam McVey
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G1-TRAM
ISPARTOF	G1A-TRAM

9

TITLE	G-LOANING
IDENTIFIER	NEWSATSIXCEN310519G-LOANING
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	A man admitted killing his niece by hitting her with his car whilst driving over the speed limit.
SUBJECT	Traffic crime; family
SUBJECT	High court; Motherwell
CREATOR	Gordon Chree
CREATOR ROLE	Commentator

10

TITLE	G-JOHNSTON
IDENTIFIER	NEWSATSIXCEN310519G-JOHNSTON
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	A 28 year old man is thought to have been stabbed near Edinburgh Castle
SUBJECT	Homicide; monument and heritage site
SUBJECT	Edinburgh Castle; Paul Smith; Balerno; Johnston Terrace
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator

11

TITLE	G-LONDON
IDENTIFIER	NEWSATSIXCEN310519G-LONDON
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	A police officer is in a coma after being struck by a car in Glasgow
SUBJECT	Road incident; police
SUBJECT	Glasgow; coma
CREATOR	Gordon Chree
CREATOR ROLE	Commentator

12

TITLE	G1-TORY
IDENTIFIER	NEWSATSIXCEN310519G1-TORY
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Conservative MP Rory Stewart says he would not back a Scottish independence referendum if he was prime minister
SUBJECT	Political candidates; ministers (government); referenda
SUBJECT	Conservative Party; Rory Stewart; Scottish independence
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
HASPART	G2-TORY

13

TITLE	G2-TORY
IDENTIFIER	NEWSATSIXCEN310519G2-TORY
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Rory Stewart says that if he was prime minister he would support Scottish issues but not back another independence referendum
SUBJECT	Political candidates; ministers (government); referenda
SUBJECT	Conservative Party; Rory Stewart; Scottish independence
CONTRIBUTOR	Rory Stewart
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G1-TORY

14

TITLE	G1-HILL
IDENTIFIER	NEWSATSIXCEN310519G1-HILL
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Conservationists are preparing to re-open Hill House in Helensburgh following extensive work to preserve it
SUBJECT	Monument and heritage site; renovation
SUBJECT	Conservation; Hill House; Charles Rennie Mackintosh; Helensburgh
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G2-HILL

15

TITLE	G2-HILL
IDENTIFIER	NEWSATSIXCEN310519G2-HILL
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Conservationists are preparing to re-open Hill House in Helensburgh following extensive work to preserve it
SUBJECT	Monument and heritage site; renovation
SUBJECT	Conservation; Hill House; Charles Rennie Mackintosh; Helensburgh
CREATOR	Susan Ripoll
CREATOR ROLE	Reporter
ISPARTOF	G1-HILL

16

TITLE	G-SPORTHAND
IDENTIFIER	NEWSATSIXCEN310519G-SPORTHAND
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Gordon hands over to Raman for the sport
SUBJECT	Sport
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G1-CELTIC
HASPART	G1A-CELTIC
HASPART	G2-CELTIC
HASPART	G3-CELTIC
HASPART	G4-CELTIC
HASPART	G5-CELTIC
HASPART	G1-REFEREE
HASPART	G2-REFEREE
HASPART	G1-UFC
HASPART	G2-UFC
HASPART	G-BACK

17

TITLE	G1-CELTIC
IDENTIFIER	NEWSATSIXCEN310519G1-CELTIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Neil Lennon becomes permanent manager of Celtic
SUBJECT	Soccer; national championship
SUBJECT	Celtic Football Club; Neil Lennon
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
HASPART	G1A-CELTIC
	G2-CELTIC
	G3-CELTIC
	G4-CELTIC
	G5-CELTIC
ISPARTOF	G-SPORTHAND

18

TITLE	G1A-CELTIC
IDENTIFIER	NEWSATSIXCEN310519G1A-CELTIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Evanna Holland reports live from Celtic Park about Neil Lennon becoming permanent manager of Celtic
SUBJECT	Soccer; national championship
SUBJECT	Football; Celtic Football Club; Neil Lennon; Scottish Cup
CREATOR	Evanna Holland
CREATOR ROLE	Presenter
HASPART	G2-CELTIC
ISPARTOF	G1-CELTIC
ISPARTOF	G-SPORTHAND

19

TITLE	G2-CELTIC
IDENTIFIER	NEWSATSIXCEN310519G2-CELTIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Neil Lennon speaks at a press conference about plans for Celtic to win a fourth treble and recruitment in the summer transfer market
SUBJECT	Soccer; national championship
SUBJECT	Celtic Football Club; Neil Lennon; Scottish Cup; recruitment; Champions League; league titles
CREATOR	Evanna Holland
CREATOR ROLE	Interviewer
CONTRIBUTOR	Neil Lennon
CONTRIBUTOR ROLE	Interviewee
ISPARTOF	G-SPORTHAND
ISPARTOF	G1-CELTIC
ISPARTOF	G1A-CELTIC

20

TITLE	G3-CELTIC
IDENTIFIER	NEWSATSIXCEN310519G3-CELTIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Evanna Holland reports live from Celtic Park about the summer transfer market
SUBJECT	Soccer; transfer
	Celtic Football Club; Neil Lennon; Peter Lawwell; recruitment
CREATOR	Evanna Holland
CREATOR ROLE	Reporter
HASPART	G4-CELTIC
ISPARTOF	G-SPORTHAND
ISPARTOF	G1-CELTIC

21

TITLE	G4-CELTIC
IDENTIFIER	NEWSATSIXCEN310519G4-CELTIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Peter Lawwell speaks at a press conference about Celtic's plans for the summer transfer market
SUBJECT	Soccer; transfer
SUBJECT	Celtic Football Club; recruitment
CONTRIBUTOR	Peter Lawwell
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G-SPORTHAND
ISPARTOF	G1-CELTIC
ISPARTOF	G3-CELTIC

22

TITLE	G5-CELTIC
IDENTIFIER	NEWSATSIXCEN310519G5-CELTIC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Evanna Holland reports live from Celtic Park about Celtic's return to pre-season training and Champions League qualifiers
SUBJECT	Soccer; national games
SUBJECT	Celtic Football Club; Champions League
CREATOR	Evanna Holland
CREATOR ROLE	Reporter
ISPARTOF	G-SPORTHAND
ISPARTOF	G1-CELTIC

23

TITLE	G1-REFEREE
IDENTIFIER	NEWSATSIXCEN310519G1-REFEREE
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Referee Kylie Cockburn is in Paris ahead of the women’s World Cup.
SUBJECT	Soccer; world cup
SUBJECT	Shelley Kerr; Kylie Cockburn
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
HASPART	G2-REFEREE
ISPARTOF	G-SPORTHAND

24

TITLE	G2-REFEREE
IDENTIFIER	NEWSATSIXCEN310519G2-REFEREE
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Referee Kylie Cockburn is interviewed about how she was selected to be a referee at the women’s World Cup, and her role as a community police offer
SUBJECT	Soccer; world cup; accomplishment; police
SUBJECT	Referee; assistant referee
CREATOR	Ronnie Charters
CREATOR ROLE	Interviewer
CONTRIBUTOR	Kylie Cockburn
CONTRIBUTOR ROLE	Interviewee
ISPARTOF	G-SPORTHAND
ISPARTOF	G1-REFEREE

25

TITLE	G1-UFC
IDENTIFIER	NEWSATSIXCEN310519G1-UFC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Stevie Ray prepares to fly to Stockholm to take on Leonardo Santos in mixed martial arts
SUBJECT	Mixed martial arts; international games
SUBJECT	Stevie Ray; Leonardo Santos
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
HASPART	G2-UFC
ISPARTOF	G-SPORTHAND

26

TITLE	G2-UFC
IDENTIFIER	NEWSATSIXCEN310519G2-UFC
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Stevie Ray speaks about taking on Leonardo Santos in mixed martial arts, who has not fought in over two and a half years
SUBJECT	Mixed martial arts; international games
SUBJECT	Stevie Ray; Leonardo Santos
CONTRIBUTOR	Stevie Ray
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G-SPORTHAND
ISPARTOF	G1-UFC

27

TITLE	G-BACK
IDENTIFIER	NEWSATSIXCEN310519
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Raman closes the sport section
SUBJECT	Sport
SUBJECT	Goodbye
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
ISPARTOF	G-SPORTHAND

28

TITLE	G1-WEATHERLINK
IDENTIFIER	NEWSATSIXCEN310519G1-WEATHERLINK
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	Kelly-Ann Woodland and Gordon Chree talk to weatherman Sean Batty about his new programme, Sean’s Scotland
SUBJECT	Weather; weather forecast; travel
SUBJECT	Promo; Sean’s Scotland; Sean Batty; Scotland; Wester Ross; Colonsay; Highlands and Islands; Deeside; Mull
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
CONTRIBUTOR	Sean Batty
CONTRIBUTOR ROLE	Reporter
HASPART	G2-WEATHER

29

TITLE	G2-WEATHER
IDENTIFIER	NEWSATSIXCEN310519G2-WEATHER
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Sean Batty presents the evening's weather for Glasgow
SUBJECT	Weather forecast
SUBJECT	Lomond Games; Helensburgh
CREATOR	Sean Batty
CREATOR ROLE	Reporter
ISPARTOF	G1-WEATHERLINK

30

TITLE	G1-MOAT
IDENTIFIER	NEWSATSIXCEN310519G1-MOAT
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	The house which inspired Peter Pan was on the brink of demolition and has been transformed into a literary centre for children
SUBJECT	Library and museum; literature; leisure venue
SUBJECT	Moat Brae; Peter Pan; JM Barrie
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator
HASPART	G2-MOAT

31

TITLE	G2-MOAT
IDENTIFIER	NEWSATSIXCEN310519
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow; Edinburgh
DESCRIPTION	The house which inspired Peter Pan was on the brink of demolition and has been transformed into a literary centre for children
SUBJECT	Library and museum; literature; leisure venue
SUBJECT	Moat Brae; Dumfries; demolition; Peter Pan; JM Barrie
CREATOR	Laura Piper
CREATOR ROLE	Reporter
CONTRIBUTOR	Joanna Lumley
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G1-MOAT

32

TITLE	G-BYE
IDENTIFIER	NEWSATSIXCEN310519G-BYE
DATE	31-05-2019
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Kelly-Ann Woodland and Gordon Chree end the evening's programme
SUBJECT	News media
SUBJECT	Goodbye
CREATOR	Kelly-Ann Woodland
CREATOR ROLE	Commentator
CREATOR	Gordon Chree
CREATOR ROLE	Commentator

Metadata records for 2014 dataset (Items 1-40)

1

TITLE	G-OPENERS
IDENTIFIER	NEWSATSIXGLA020614G-OPENERS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	John MacKay introduces the news headlines
SUBJECT	Referenda; motor car racing; road accident and incident; soccer; war
SUBJECT	Scottish independence; Celtic Football Club; Roy Keane; Steve Clarke; WW1; First World War; Great War
CREATOR	John MacKay
CREATOR ROLE	Commentator

2

TITLE	G1-POLL
IDENTIFIER	NEWSATSIXGLA020614G1-POLL
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Election campaign gets underway for Scottish independence,
SUBJECT	Referenda; political campaigns; political parties and movements
SUBJECT	Scottish independence; undecided voters; Yes campaign; Better Together
CREATOR	John Mackay
CREATOR ROLE	Commentator
HASPART	G2-POLL

3

TITLE	G2-POLL
IDENTIFIER	NEWSATSIXGLA020614G2-POLL
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	A poll reveals that men are equally split on voting yes or no to Scottish independence but fewer women are in support of independence
SUBJECT	Referenda; political campaigns; political parties and movements; demographics; gender
SUBJECT	Scottish independence; undecided voters; Yes campaign; Better Together; Alex Salmond; SNP
CREATOR	Claire Stewart
CREATOR ROLE	Reporter
ISPARTOF	G1-POLL

4

TITLE	G1-TORIES
IDENTIFIER	NEWSATSIXGLA020614G1-TORIES
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Political parties in Scotland want responsibility to set income tax
SUBJECT	Economic policy; taxation; referenda
SUBJECT	Scottish government; income tax; devolution; Strathclyde Commission; Holyrood
CREATOR	John Mackay
CREATOR ROLE	Commentator
HASPART	G2-TORIES
HASPART	G3-TORIES

5

TITLE	G2-TORIES
IDENTIFIER	NEWSATSIXGLA020614G2-TORIES
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	The three unionist parties have made their pitch on more powers for Holyrood
SUBJECT	Economic policy; taxation; referenda
SUBJECT	Scottish government; income tax; devolution; Strathclyde Commission; Holyrood; Better Together; Yes campaign; Conservative Party; Margaret Thatcher; Ted Heath; Alec Douglas-Home; unionism
CREATOR ROLE	Bernard Ponsonby
CONTRIBUTOR	Reporter
HASPART	G3-TORIES
ISPARTOF	G1-TORIES

6

TITLE	G3-TORIES
IDENTIFIER	NEWSATSIXGLA020614G3-TORIES
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Bernard Ponsonby reports live on the three unionist parties' proposals on more powers for Holyrood
SUBJECT	Election; referenda; government policy
SUBJECT	2015 general election; unionist; Conservative Party; Labour Party; Liberal Democrats; Ed Miliband; Nick Clegg; Better Together; Yes campaign; Scottish independence
CREATOR	John Mackay
CREATOR ROLE	Interviewer
CONTRIBUTOR	Bernard Ponsonby
CONTRIBUTOR ROLE	Interviewee
ISPARTOF	G1-TORIES
ISPARTOF	G2-TORIES

7

TITLE	G-TONIGHT
IDENTIFIER	NEWSATSIXGLA020614G-TONIGHT
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Promo for Scotland Tonight with guest Ruth Davidson
SUBJECT	Politics; news media
SUBJECT	Scotland Tonight; Scottish Conservative Party; Ruth Davidson
CREATOR	John Mackay
CREATOR ROLE	Commentator

8

TITLE	G1-RALLY
IDENTIFIER	NEWSATSIXGLA020614G1-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Scotland's top law officer has refused to rule out a criminal prosecution over death of spectators at a car rally in the Borders
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Scottish Borders
CREATOR	John Mackay
CREATOR ROLE	Commentator
HASPART	G2-RALLY
	G3-RALLY
	G4-RALLY
	G5-RALLY
	G-WEBSITE

9

TITLE	G2-RALLY
IDENTIFIER	NEWSATSIXGLA020614G2-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Close friends and neighbours speak about those who died at the car rally tragedy in the Scottish Borders
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Scottish Borders
CREATOR	Sharon Frew
CREATOR ROLE	Reporter
ISPARTOF	G1-RALLY

10

TITLE	G3-RALLY
IDENTIFIER	NEWSATSIXGLA020614G3-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Gordon Chree reports live from Kelso about the investigation into the car rally tragedy in the Scottish Borders
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Scottish Borders
CREATOR	John Mackay
CREATOR ROLE	Interviewer
CONTRIBUTOR	Gordon Chree
CONTRIBUTOR ROLE	Interviewee
ISPARTOF	G1-RALLY

11

TITLE	G4-RALLY
IDENTIFIER	NEWSATSIXGLA020614G4-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Gordon Chree interviews Scotland’s top law officer about the possibility of criminal prosecution in the car rally tragedy case
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Scottish Borders
CREATOR	Gordon Chree
CREATOR ROLE	Interviewer
CONTRIBUTOR	Frank Mulholland
CONTRIBUTOR ROLE	Interviewee
ISPARTOF	G1-RALLY
ISPARTOF	G3-RALLY

12

TITLE	G5-RALLY
IDENTIFIER	NEWSATSIXGLA020614G5-RALLY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Justice secretary Kenny MacAskill will give a statement to parliament about the car rally tragedy in the Borders
SUBJECT	Road accident and incident; motor car racing; punishment (criminal)
SUBJECT	Justice secretary; Commonwealth Games; Ryder Cup; Scottish Border
CREATOR	Gordon Chree
CREATOR ROLE	Reporter
ISPARTOF	G1-RALLY
ISPARTOF	G3-RALLY

13

TITLE	G-WEBSITE
IDENTIFIER	NEWSATSIXGLA020614
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Promo for website
SUBJECT	News media
SUBJECT	Promo
CREATOR	John Mackay
CREATOR ROLE	Commentator
ISPARTOF	G1-RALLY

14

TITLE	G-DEMENTIA
IDENTIFIER	NEWSATSIXGLA020614G-DEMENTIA
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Levels of care for dementia sufferers have been criticised
SUBJECT	Medical conditions; health facility
SUBJECT	Dementia; NHS; human rights
CREATOR	John MacKay
CREATOR ROLE	Commentator

15

TITLE	G-ASHRAF
IDENTIFIER	NEWSATSIXGLA020614G-ASHRAF
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	A judge told Mohammed Ashraf that if he had not pled guilty when he did, the sentence would have been six years
SUBJECT	Sex crime
SUBJECT	Mohammed Ashraf; taxi driver
CREATOR	John Mackay
CREATOR ROLE	Commentator

16

TITLE	G-TRAMS
IDENTIFIER	NEWSATSIXGLA020614G-TRAMS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	The opening of Edinburgh's tram network
SUBJECT	Road transport; commuting
SUBJECT	Trams; Edinburgh tram project
CREATOR	John Mackay
CREATOR ROLE	Commentator

17

TITLE	G1-CITY
IDENTIFIER	NEWSATSIXGLA020614G1-CITY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	The launch of new channel STV Glasgow
SUBJECT	News media
SUBJECT	Promo; STV Glasgow; TV drama; magazine programming
CREATOR	John MacKay
CREATOR ROLE	Commentator
HASPART	G2-CITY

18

TITLE	G2-CITY
IDENTIFIER	NEWSATSIXGLA020614G2-CITY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	The launch of new channel STV Glasgow
SUBJECT	News media; television; communities
SUBJECT	Promo; STV Glasgow; TV drama; magazine programming; Glasgow; local news
CREATOR	Lucy Whyte
CREATOR ROLE	Reporter
ISPARTOF	G1-CITY

19

TITLE	G-TONER
IDENTIFIER	NEWSATSIXGLA020614G-TONER
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	A man has been arrested in connection with a death that happened in 2004
SUBJECT	Homicide
SUBJECT	Martin Toner; Paisley Sheriff Court
CREATOR	John MacKay
CREATOR ROLE	Commentator

20

TITLE	G-DOYLE
IDENTIFIER	NEWSATSIXGLA020614G-DOYLE
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	DNA matching the man accused of murdering Elaine Doyle was found on her naked body but not her clothes
SUBJECT	Homicide; trial (court)
SUBJECT	DNA; Elaine Doyle; John Docherty
CREATOR	John MacKay
CREATOR ROLE	Commentator

21

TITLE	G-DIXON
IDENTIFIER	NEWSATSIXGLA020614G-DIXON
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Police are looking for more information about a sexual assault that took place in Govanhill
SUBJECT	Sex crime
SUBJECT	Glasgow; Govanhill
CREATOR	John MacKay
CREATOR ROLE	Commentator

22

TITLE	G-COMINGUP
IDENTIFIER	NEWSATSIXGLA020614G-COMINGUP
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Promo for a report on seven brothers who fought in WW1
SUBJECT	War
SUBJECT	Promo; WW1; First World War; Great War
CREATOR	John MacKay
CREATOR ROLE	Commentator

23

TITLE	G2-WEATHERPRO
IDENTIFIER	NEWSATSIXGLA020614G2-WEATHERPRO
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Promo for weather forecast
SUBJECT	Weather; weather forecast
SUBJECT	Promo
CREATOR	Sean Batty
CREATOR ROLE	Reporter

24

TITLE	G-SPORTHAND
IDENTIFIER	NEWSATSIXGLA020614G-SPORTHAND
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	John MacKay introduces the sport
SUBJECT	Soccer
SUBJECT	Celtic Football Club
CREATOR	John MacKay
CREATOR ROLE	Commentator
HASPART	G1-KEANE

25

TITLE	G1-KEANE
IDENTIFIER	NEWSATSIXGLA020614G1-KEANE
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Roy Keane will not become the new Celtic manager
SUBJECT	Soccer
SUBJECT	Celtic Football Club; Celtic; Celtic FC; Republic of Ireland national football team; Roy Keane
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
HASPART	G1A-KEANE
ISPARTOF	G-SPORTHAND

26

TITLE	G1A-KEANE
IDENTIFIER	NEWSATSIXGLA020614
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Roy Keane will not become the new Celtic manager
SUBJECT	Soccer
SUBJECT	Celtic Football Club; Celtic; Celtic FC; Republic of Ireland national football team; Roy Keane; Martin O'Neill
CREATOR	Grant Russell
CREATOR ROLE	Reporter
HASPART	G2-KEANE
ISPARTOF	G1-KEANE

27

TITLE	G2-KEANE
IDENTIFIER	NEWSATSIXGLA020614G2-KEANE
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Fans react to the news that Roy Keane will not become the new Celtic manager
SUBJECT	Soccer
SUBJECT	Celtic Football Club; Celtic; Celtic FC; Republic of Ireland national football team; Roy Keane; Malky Mackay
ISPARTOF	G1A-KEANE

28

TITLE	G3-KEANE
IDENTIFIER	NEWSATSIXGLA020614G3-KEANE
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Grant Russell gives names of potential candidates to be the new Celtic manager
SUBJECT	Soccer
SUBJECT	Celtic Football Club; Celtic; Celtic FC; Peter Lawwell; Roy Keane; Steve Clarke; Malky Mackay
CREATOR	Raman Bharwaj
CREATOR ROLE	Interviewer
CONTRIBUTOR	Grant Russell
CONTRIBUTOR ROLE	Interviewee
ISPARTOF	G1-KEANE

29

TITLE	G1-TENNIS
IDENTIFIER	NEWSATSIXGLA020614G1-TENNIS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Andy Murray is edging closer to a place in the quarterfinals of the French Open
SUBJECT	Tennis; international games
SUBJECT	Andy Murray; Fernando Verdasco; French Open
CREATOR	John MacKay
CREATOR ROLE	Commentator

30

TITLE	G1-RUGBY
IDENTIFIER	NEWSATSIXGLA020614G1-RUGBY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Al Kellock says his international career is far from over, despite being left out of the Scotland squad for the summer tour.
SUBJECT	Rugby
SUBJECT	Al Kellock; Glasgow Warriors; Leinster; Pro12
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
HASPART	G2-RUGBY

31

TITLE	G2-RUGBY
IDENTIFIER	NEWSATSIXGLA020614G2-RUGBY
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Al Kellock speaks about his career
SUBJECT	Rugby
SUBJECT	Al Kellock; Glasgow Warriors; Leinster; Pro12
CONTRIBUTOR	Al Kellock
CONTRIBUTOR ROLE	Speaker
ISPARTOF	G1-RUGBY

32

TITLE	G-KILMARNOCK
IDENTIFIER	NEWSATSIXGLA020614G-KILMARNOCK
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Jamie Hamill has signed a three-year contract with Kilmarnock Football Club
SUBJECT	Soccer
SUBJECT	Kilmarnock; Kilmarnock Football Club; Kilmarnock FC; Killie; Jamie Hamill; Ann Budge
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator

33

TITLE	G-FALKIRK
IDENTIFIER	NEWSATSIXGLA020614G-FALKIRK
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Falkirk are searching for a new head coach after Gary Holt quit
SUBJECT	Soccer
SUBJECT	Falkirk; Falkirk Football Club; Falkirk FC; Gary Holt; Scottish Premiership; Neil Adams
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator

34

TITLE	G-GOLF
IDENTIFIER	NEWSATSIXGLA020614G-GOLF
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Stephen Gallacher narrowly missed out on tour victory
SUBJECT	Golf
SUBJECT	Stephen Gallacher; Nordea Masters; Tonghcai Jaidee
CREATOR	Raman Bhardwaj

35

TITLE	G-BACK
IDENTIFIER	NEWSATSIXGLA020614G-BACK
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Raman closes the sport for the programme
SUBJECT	Sport; news media
SUBJECT	Promo
CREATOR	Raman Bhardwaj
CREATOR ROLE	Commentator
CONTRIBUTOR	John MacKay
CONTRIBUTOR ROLE	Speaker

36

TITLE	G1-WEATHERLINK
IDENTIFIER	NEWSATSIXGLA020614G1-WEATHERLINK
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	John introduces the weather forecast
SUBJECT	Weather; weather forecast
CREATOR	John MacKay
CREATOR ROLE	Commentator
HASPART	G2-WEATHER01
ISPARTOF	G-WEATHERPRO

37

TITLE	G2-WEATHER01
IDENTIFIER	NEWSATSIXGLA020614G2-WEATHER01
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Sean Batty presents the evening's weather
SUBJECT	Weather forecast
CREATOR	Sean Batty
CREATOR ROLE	Reporter
ISPARTOF	G1-WEATHERLINK
ISPARTOF	G2-WEATHERPRO

38

TITLE	G1-BROTHERS
IDENTIFIER	NEWSATSIXGLA020614G1-BROTHERS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Descendants of brothers who fought in WWI commemorate their sacrifices
SUBJECT	War; ceremony
SUBJECT	WWI; World War 1; First World War; Great War
CREATOR	John Mackay
CREATOR ROLE	Commentator
HASPART	G2-BROTHERS

39

TITLE	G2-BROTHERS
IDENTIFIER	NEWSATSIXGLA020614G2-BROTHERS
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	Descendants of brothers who fought in WWI commemorate their sacrifices
SUBJECT	War; ceremony
SUBJECT	WWI; World War 1; First World War; Great War; commemoration; centenary
CREATOR	Jennifer Harrild
CREATOR ROLE	Reporter
ISPARTOF	G1-BROTHERS

40

TITLE	G-BYE
IDENTIFIER	NEWSATSIXGLA020614G-BYE
DATE	02-06-2014
EDITION	STV News at Six
REGION	Glasgow
DESCRIPTION	John MacKay closes the programme
SUBJECT	News media
SUBJECT	Promo; goodbye
CREATOR	John MacKay
CREATOR ROLE	Commentator