UDC at the BBC

Fran Alexander, Kathryn Stickley, Vicky Buser, and Libby Miller
United Kingdom

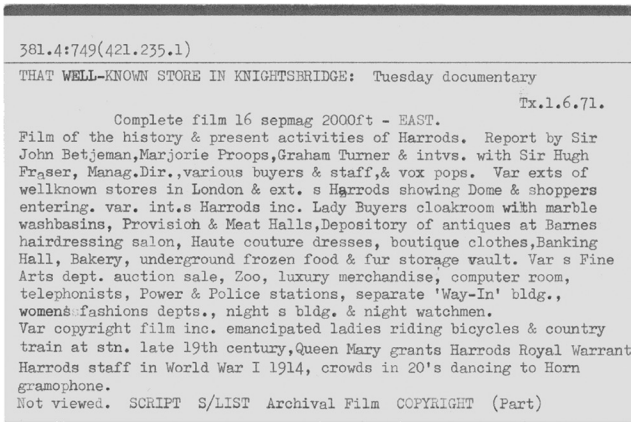
The BBC Archive

The BBC Archive (http://www.bbc.co.uk/archive/) is one of the world's largest multimedia archives, held in 27 locations across the UK. The Archive contains over 2 million items of TV and video, 300,000 hours of audio, 6 million still photographs, over 4 million items of sheet music, and over half-a-million documents and records. It is a working media library, fulfilling some 4,000 loans per week, as well as preserving content as part of the UK's national cultural heritage. A team of cataloguers and media managers classify a selection of current content, as well as enhancing cataloguing and classification of legacy content.

There are two major classification schemes used in the Archive, both numerical, and one based on UDC. Lonclass, based on UDC, was developed first, then Telclass, which is used by the Natural History Unit in Bristol. In addition, there are many and various controlled vocabularies that have been developed to tag content in the different nations (Scotland, Wales, and Northern Ireland) and the English regions.

The History of the Archive Classifications

The BBC's early film archives, which date back to 1936, related primarily to news. The cataloguing was typed onto strips of card that were then slotted into wall-mounted metal frames and ordered alphabetically by keywords. A 5-volume thesaurus listing all the keywords and synonyms was used as a guide to the correct "approved" keyword. This was a cumbersome but fragile system requiring a lot of wall space. It was certainly not user friendly, as the cardboard strips themselves often ended up on the floor, meaning that they needed to be re-sorted and re-inserted into the frames by hand. Inevitably over the years some of the cataloguing strips were lost altogether.



381.4:749(421.235.1)
THAT WELL-KNOWN STORE IN KNIGHTSBRIDGE: Tuesday documentary
Tx.1.6.71.
Complete film 16 sepmag 2000ft - EAST.
Film of the history & present activities of Harrods. Report by Sir John Betjeman,Marjorie Proops,Graham Turner & intvs. with Sir Hugh Fraser, Manag.Dir.,various buyers & staff,& vox pops. Var exts of wellknown stores in London & ext. s Harrods showing Dome & shoppers entering. var. int.s Harrods inc. Lady Buyers cloakroom with marble washbasins, Provision & Meat Halls,Depository of antiques at Barnes hairdressing salon, Haute couture dresses, boutique clothes,Banking Hall, Bakery, underground frozen food & fur storage vault. Var s Fine Arts dept. auction sale, Zoo, luxury merchandise, computer room, telephonists, Power & Police stations, separate 'Way-In' bldg., womens fashions depts., night s bldg. & night watchmen.
Var copyright film inc. emancipated ladies riding bicycles & country train at stn. late 19th century,Queen Mary grants Harrods Royal Warrant Harrods staff in World War I 1914, crowds in 20's dancing to Horn gramophone.
Not viewed. SCRIPT S/LIST Archival Film COPYRIGHT (Part)

Figure 1 *- A cataloguing card from 1971. The top of the card has a red flash (stripe) indicating that it was colour film, the yellow card means a programme (as opposed to green cards for news items, etc.). A black flash indicated VT (videotape). This sort of metadata was transformed into distinct database fields in Infax. Such cards were hand-typed until the late 1970s.*

Lonclass was introduced in 1964, as the intrinsic value of media assets gained recognition and TV programmes, such as documentaries, began to be catalogued with the aim of helping other programme makers find and re-use relevant content. Sturdier 6" x 4" cataloguing cards replaced the flimsy strips, with the classification numbers starting at the top-left hand corner, with a rule below the number, to separate it from the programme title and transmission date.

Lonclass uses UDC at its core, but was created "bottom up", with new terms, including pre-coordinated terms, added purely on warrant. Whilst that resulted in not all of the available UDC terms being on offer, it meant that researchers could be confident that they would always get a return on their searches. Lonclass currently has only 20,000 core "simple" concepts, but 330,000 complex concepts. In an attempt to maintain consistency, individual terms were given rules with guidance on correct application. UDC was extended greatly to enhance the usefulness of Lonclass for cataloguing audio-visual content. New non-reversible auxiliaries were added such as Vehicular Motion (e.g. M57 for take off) and Camera Motion (e.g. R13 for ground to air shots).

By the late 1970s, the size and sheer weight of the card catalogue and its index had become a concern. The subject index was computerised in 1981, making it easier to create new terms, and in 1984 the Broadcast Archive began cataloguing using an in-house computer system – Infax – which was written in Informix and ultimately combined a subject index, subject catalogue, stock management, and loans system. In 1985 the Archive took on the responsibility for cataloguing the majority of BBC Network TV output (the programmes broadcast across all the nations and regions).

A long data migration exercise was undertaken to add the various card catalogues and the strips catalogue to Infax. Other systems and BBC areas joined Infax, such as Radio and the Music library. As the catalogue continued to grow, the need for more accurate and specific indexing grew and the subject index grew along with it, at times unrestrained.

In 1979 another classification scheme was established, Telclass, which has been used by the Natural History Unit based at BBC Bristol since 1982. The aim was to create a new, independent classification scheme designed specifically for television content. Telclass, which is not a UDC-based classification, was eventually made accessible through Infax, although most cataloguers only worked with one classification or the other.

To the untrained user, Infax could be confusing, but further advances in technology opened up subject searching to even the most inexperienced researcher – the introduction of automated translations into natural language of the preferred terms for Lonclass. This ensured that searchers did not need to look up the classification number in a separate index, it meant a version of Infax was accessible through a standard web browser with a more user-friendly interface, and it also facilitated free text searching (which included searching the terms in the natural language translations of the classification numbers). Whilst not being perfect (the free text search did not include synonyms, the ability to search Lonclass subjects as natural language meant that non-librarians could use the taxonomy without even realising it.

Classifications for the future

The BBC is undertaking a major project to create an integrated digital production and archival system for new content. This represents a major change in the place of the Archive, setting it at the heart of production processes. Archive content (most of which has not yet been digitised) is no longer seen as just for historical supplements, but as a source of inspiration for new productions and ideas. As the Archive now approaches this new stage in its life, Lonclass will again be adapted and migrated to aid the search and re-use of the BBC's archive content by a wider range of non-specialist users.

With advances in technology, metadata can be harvested far more easily from sources such as digital cameras and from the electronic documentation created by programme makers. Meanwhile, production teams have their own requirements to find digital content while they are making programmes, so working out how to harvest their often folksonomic logging and adapt it to serve archival needs is another challenge. However, if such metadata can be preserved and used, the benefits include avoiding having to log the same content twice and capturing the specialist knowledge of the production teams – the precise location of a particular shot, for example.

Bringing such metadata together into a useful taxonomy has been a major driver of how Lonclass and Telclass should be adapted for future use. Attempting to return Lonclass to its UDC roots would be a labour intensive process, as this would involve unpicking the areas where UDC syntactical rules were not followed, sometimes on an ad hoc basis. However, the UDC heritage will be preserved in the new taxonomy, which will retain a complex faceted structure and editorial policy will continue to mirror UDC principles. The legacy classifications will be preserved and will remain a route for finding archival content, as well as a source of valuable information on past programmes.

A new purpose for Lonclass?

BBC Research and Development is currently contributing to NoTube (http://www.notube.tv/), a collaborative research project about the convergence of Web and TV funded by the EU. Part of the project involves investigating how the merging of the Semantic Web and TV using metadata can enhance the TV viewing experience. One aspect is making it easier for people to choose what to watch by serendipitous browsing of large video collections, so ways to help users navigate through such collections by following interesting connections between programmes are being investigated, including how Lonclass could be used.

Navigating large collections and particularly choosing what to watch is a central unsolved problem for on-demand video, as the user is faced with a vast number of possibilities. A recent NoTube demo (http://www.flickr.com/photos/nicecupoftea/4945125947/) shows a Web-based user interface for navigating a substantial subset of three years' worth of programmes in the BBC Archive (based on a private BBC research project). From any programme, the user can follow suggestions for similar programmes based on the number of common Lonclass links between them. The idea is to support "hours of fascinated clicking" through the programme archive, similar to the way that following links in Wikipedia articles can take users on surprising and unexpected journeys through the content.

This technique allows for programmes of interest in the long-tail of the BBC Archive to surface, whereas recommendation techniques based on collaborative filtering (*"You might like Animal Park, because viewers who watched Animal Park also watched Springwatch"*) rely on other people having watched something before it can be suggested. The technique can also generate a more meaningful explanation about why a programme has been suggested (*e.g. this programme is also about medical research and the history of inventions*) to help users make more accurate decisions about whether to watch something or not.

Early NoTube experiments with linked data programme recommendations were based on automated entity recognition and extraction of terms in programme metadata being matched to relevant DBpedia concepts. The results suggested that using DBpedia alone produces recommendations that are too general, such as *"Eastenders is recommended because you like programmes made in the 80s"*. One conclusion was that there are inherent limitations to such a scattergun approach and it was proposed that using Lonclass might produce more useful results, given its diversity combined with its TV-focus. A comparison of Lonclass and DBpedia also suggested that Lonclass might be better for abstract concepts, for describing the "aboutness" of a programme. Further, since a human indexer had assigned the specific link between the programme and the Lonclass term, the recommendations might be more accurate as well as more interesting.

User testing of these theories are being planned and a potential by-product of the testing might be the collection of enough data to reverse engineer the links or paths that might exist between the programmes people like, based on Lonclass terms. If successful, this could provide interesting insights into the *types of links* between programmes that are interesting to people – for example, are links based on location more interesting than links based on time period, or is a combination of location *and* time period more interesting still? Perhaps it might then be possible to organise people into categories based on their preferences about programmes and to make useful programme suggestions based on these categories?

Another topic the NoTube team are investigating is how to make it easier for people to discover more about a programme they are watching. Initial experiments suggest that DBpedia links can automatically provide quirky and interesting links to more information from the Web about a specific programme. Connecting Lonclass with DBpedia would provide additional input for such *"Do you want to know more?"* questions. Linking Lonclass to the Linked Data cloud would allow for interesting new connections to be made across the Web, creating additional background knowledge and context about a programme that could be pushed to the viewer using a second screen such as an iPad.

The NoTube experiments with Lonclass demonstrate how it could be re-used in new and potentially useful ways that address realistic user scenarios. The demo shows how such a UDC-based traditional classification system can be adapted to create rich Web-based user journeys based on serendipitous browsing and content discovery. In this way the value of Lonclass can be exposed to new audiences, extending its usage beyond the domain of the "expert" to anyone, without the need for any specialist knowledge of cataloguing or classification schemes.

During the remainder of the NoTube project, which ends in February 2012, the team will continue to investigate ways of unlocking the potential of Lonclass for new types of users in ways that enhance the TV experience.

Conclusion

The BBC recognises that it cannot run an effective archive without robust methods of classifying its content. Free text searching on its own cannot fulfil the very specific requirements of our researchers. Projects such as NoTube illustrate how useful and adaptable classification schemes such as Lonclass and Telclass continue to be. Without classification, the Archive would be impossible to navigate and the valuable content could no longer be located.