IFLA WLIC 2013
SINGAPORE

Submitted on: August 6, 2013

Caxton to Cloud: The Future of news preservation, storage and access

Patrick Fleming
Head of Operations, British Library
London, United Kingdom
E-mail: Patrick.fleming@bl.uk

Abstract:

This paper outlines how the British Library is meeting the challenge of future news preservation, storage and access. It tracks the transition from traditional hard copy collections of newspapers in a collection with content dating back from the 1800's to today through to the current transformational changes taking place with the creation of a modern, sophisticated low oxygen physical storage environment. It also outlines the collection challenges involved in migrating from a core newspaper collection to a news content strategy which captures and makes available multimedia formats.

Way back in 1473 William Caxton created news when he published a book from an English printing press.

It was a historic breakthrough and stimulated a new culture of collecting.

With collecting came the challenge of preservation, storage and access and with that came a relentless change in the volume of publishing and varying formats.

For centuries the challenge for libraries worldwide was to deal predominantly with the physical. That changed with the arrival of radio and television but collecting from these mediums stayed largely outside the remit of most libraries.

Then came the Internet and the revolution of now, any time, any place any where media ns nowhere is this reflected more radically than in the production and dissemination of 'news.

This paper illustrates how the British Library is dealing with the challenge now and into the future of news preservation and access.

The purpose of preservation is to ensure protection of information of enduring value for access by present and future generations.

Libraries and archives have served as the central institutional focus for preservation, and both types of institutions include preservation as one of its core functions.

In recent decades, many major libraries and archives have established formal preservation programs for traditional materials which include regular allocation of resources for preservation, preventive measures to arrest deterioration of materials, remedial measures to restore the usability of selected materials, and the incorporation of preservation needs and requirements into overall program planning.

Preservationists within the library and archival community have been instrumental in developing an array of tools and methodologies to reduce the decay of traditional materials and to restore print content that has deteriorated to such an extent that their longevity and usability are threatened. Provisions for fire protection and adequate environmental controls frequently are incorporated into new library and archival facilities. Rehousing of acid-based paper materials is a common task in many repositories and microfilming is used extensively and cost effectively to preserve endangered materials.

The British Library knows this challenge well and the challenges continue to increase.

As well as storing rare, valuable and vulnerable print collections in our basement facilities at St Pancras we have further storage facilities at Boston Spa in Yorkshire designed to provide the optimum storage preservation standards for physical material.

At Boston Spa we have the £26m Additional Storage Building. The fully-automated facility comprises 262 linear km of extra shelf space (enough to stretch from London to Manchester) for the collection which is currently expanding at a rate of 12.5km of linear shelf space per year.

At present this building plays a key role in assisting the Library deliver its current Newspaper Strategy, housing newspaper content whilst we complete a new storage building for newspapers.

For many years right up to the close of the 20th century, the Library, like other libraries worldwide considered news preservation, storage and access as broadly a print and microfilm based issue but the world is changing rapidly bringing with it enormous challenges for institutions seeking to capture the outputs from the world of news.

For the British Library news collection had originally begun with the British Museum in the 1820's but it ran out of storage space by the turn of the 20th century. Only widespread opposition stopped a proposal to dispose of all the provincial newspapers collected and a storage building was constructed at Colindale opening in 1905.

The building was home to a large collection of English local newspapers, Welsh, Scottish and Irish titles, but London newspapers and foreign copies were kept at the main British Museum building in Bloomsbury with a weekly transport system - initially consisting of a horse-drawn cart - delivering papers to readers in the Newspaper Reading Room at the museum. The repository was full after 20 years and

in 1928 a purpose-built newspaper library was commissioned at Colindale which was completed four years later. Since then the newspaper collections has attracted scholars from all over the world.

Today some 30,000 researchers consult documents from the national newspaper collection each year.

- 56 per cent of material viewed is from local newspapers from the UK and Ireland
- 31 per cent view national titles
- 43 per cent are pursuing personal interests, such as trying to piece together their family history
- A further 29 per cent are using the resources to research school or university projects

Home to 750 million pages of newspapers and quantities of microfilm the collection was at great risk. The building was inadequate, the storage facilities poor and we were running out of space.

In 2007 we created a Newspaper strategy for the long-term storage, preservation and access to the collection. With funding of £33m from Government we are on the last leg of a long journey that will see the physical collections arrive in a new home in Yorkshire and will be steeled in by the end of 2014.

Colindale and the Newspaper Library will close for the last time on November 8 this year. It is the first time the full collection will be stored outside of London.

The new state-of-the-art facility will be controlled by some 160km of robotic shelf space, with newspapers kept in low oxygen to prevent risk of fire.

The new building features full temperature and humidity control to maintain collection items in archival standard conditions and drastically increase the life span of the physical items.

However, that is only part of the British Library's news story.

Since we embarked on our Newspaper Strategy in 2007 the world of news has changed beyond measure and will continue to do so.

Newspapers worldwide have experienced ground shifting decline as readers migrate to other sources, television, radio, the web and social media or a mix of all to satisfy their news appetite.

Publishers have sought to meet this change to news access promiscuity by providing news in multiple formats. Where once there was a newspaper, there now is a newspaper plus, a news website, a news radio station, a news television outlet, news on the web, news on tablet, news on mobiles and news archives or a mix of some of these.

For collecting institutions this poses enormous problems and requires a step change in attitudes.

The British Library has met these challenges head on and has created a new collection strategy for news.

This strategy is a plan of action leading from the recommendations made in the News Content Review in January 2013.

The Review concluded that the British Library must change from being a newspaper library to becoming a news centre, incorporating the full range of news media – newspapers, news websites, television news, radio news, and other media.

Where once we collected the entire print output of the the newspaper industry in the UK and Ireland we've concluded, not surprisingly that it is simply impossible to collect everything. So the new collecting remit is defined primarily by news that is produced in the UK, or which is receivable in the UK and has had an impact on British perceptions of what is 'news'.

The strategy is a reflection of the significant changes in news production and consumption taking place today, but it also reflects how news has always been consumed. News does not exist in any one form. It is sought out and selected by its users, from the multiple forms of information on offer. Such an understanding of news places the user at the heart of the process.

The core aims are:

- To develop the national published archive through legal deposit
- To support UK research through collecting and connecting to contemporary content
- To support research and culture through developing world-class primary research collections.

It sets the framework for the challenge the Library faces over the next 5-10 years. It stresses the need to present content in terms of disciplines and subjects, overlaid by format expertise, to have legal deposit underpin all content development activity, to continue to make the print to digital transition and increasingly to deliver content by connecting to external partners.

The news content strategy keeps within this framework, and has these key elements:

- The Library's news offering should incorporate the full range of news media – newspapers, news websites, television news, radio news, and other media – through a combination of legal deposit, purchase and voluntary deposit, capture through copyright exception, and connecting to both licensed content and content shared with strategic partners

- The Library should view news as part of the broader media landscape, finding the news content it requires by collecting or connecting to the UK media world (print, web, audiovisual), of which news forms a fundamental part
- The Library's news content should comprise primarily news most relevant to UK users, meaning news produced in the UK or which has had an impact on the UK
- News content that falls outside the definition of news produced in the UK or which has impacted upon the UK should be covered by other subject-led areas of the content strategy
- The content strategy for news media is underpinned by legal deposit collecting, both print and non-print, but incorporates audiovisual media that lie outside legal deposit
- The Library must be a champion of regional news, including regional newspapers, hyperlocal websites, community radio and regional television news
- The Library primarily collects and connects to published news, not raw news data
- The Library's news content should be made as widely available as possible to UK audiences, offering content online through licence, subscription, copyright exception and partnership arrangements, as well as maintaining physical research centres in London and Boston Spa
- The Library recognises that the concept of 'news' can be expanded to embrace anything of relevance to a particular community at a particular point in time, which long-term could have considerable impact on how it describes content and the services that it offers

The strategy is our response to significant changes in news production and consumption, and in research itself together with the Library's other relevant collecting strategies and wider core vision. These changes include:

- The possible end of newspapers in their present, paper-based form
- Media convergence and the emergence of new news media forms
- The slow death of regional newspaper publishing as advertising revenues go online, but with rise in local, alternative publishing routes
- Declining newspaper circulations
- News content rapidly going online (and mobile) without having yet reached any tipping point
- Changes in journalistic practice, including the rise in blogging and citizen journalism
- Changes in researcher habits, particularly the growth in digital scholarship

Our relevant other strategies and vision include:

- **The Library's 2020 Vision** – in particular, guaranteeing access for future generations, and enabling access to everyone who needs to do research
- **Newspaper Programme** – by providing a content pathway for the long term storage, preservation and access of the British Library's newspaper collection
- **Digital scholarship strategy** – by engaging with new and existing user communities to meet and anticipate research needs, making full use of new possibilities offered by technology
- **Web archiving strategy** – by helping ensure that the web archives are used for scholarly research in a range of disciplines, used as part of the Library's overall digital collections
- **Audio-visual strateg**y – by furthering the intention to make audio-visual media an integral part of the research experience that the British Library provides
- **Resource discovery strategy** – by contributing relevant content that meets the needs of researchers where and when they want
- **Digitisation strategy** – by contributing to the critical mass of digitised content that will open up access to the British Library's collections
- **Radio policy** – by delivering to researchers valued but previously unavailable knowledge and creative content covering a broad array of subjects
-

The successful bringing together of news forms into a single content strategy is inevitably dependent on a number of factors, several of which relate to the British Library's content management ambitions overall.

- Ongoing newspaper digitisation plans with brightsolid, our newspaper digitisation partner.
- The introduction and future success of a News and Media Centre in February 2014 as a research space and as a launchpad for innovation
- Successful service operation of the Newspaper Storage Building
- Effective capture of daily news web sources, incorporating audiovisual media
- Accurate measurement and monitoring of print to digital transition
- Effective searching and presentation of news using the Library's Explore search service
- A presentation layer able to facilitate different news media, including external content
- Partnership relationships with the British Film Institute and the BBC
- Increased recording and delivery of radio news as part of national radio archive plans
- Resource discovery that links together news media by shared time period
- Long-term digital storage and access provided by the Library's in house Digital Library System
- Parity of search results across different media, including investment in speech recognition technologies to facilitate word searching of audio and video

- Linked Open Data approach for news content, employing semantic media to enable automatic metadata generation, time-based navigation, entity extraction etc.
- Ongoing engagement with audiences of all kinds
- Curatorial support to collect, connect to and understand news at a period of critical change
- Matrix working across British Library directorates

Our aim is to be able to offer to researchers an extensive British news collection in keeping with the Library's status as a world class news resource, amassed through a combination of legal deposit, purchase and donation, capture through copyright exception; and connecting to both licensed content and content shared with strategic partners. This will be supplemented by a representative selection of news sources from around the world, chiefly by connecting to electronic resources.

It is most practical to outline the intentions for the individual news media, but these general principles apply for developing news content overall:

- Collect (and connect to) UK published national and regional news across all media
- Ensure common resource discovery
- Develop single-screen playback for all media
- Facilitate cross-linkages in time, place and theme
- Enable linkages across associated media e.g. broadcast video clips on websites
- Learn from audience usage what linkages and which resources to develop further
- Involve users in content selection and digitisation decisions
- Communicate messages that bring the news media together
- Create and provide an online viewing and editing tools enabling researchers to connect, collect and create content that is relevant to their research and interests

The news collection needs to be considered in two ways: the historical corpus, and the current and ongoing collection. Stressing news currency will be an important element of stressing the Library's position as a news centre, capturing the world's matters today while illustrating that behind every such story lies a history that the Library can help uncover.

The tables below summarise news content overall in terms of individual instances (issues, programmes or websites), for historical news and current news capture, to the end of each calendar year, with projected figures to 2017.

Current (i.e. what is collected as current news)

Medium	2013	2014	2015	2016	2017	Assumptions
Newspapers (*issues*)	1,717	1,680	1,620	1,510	1,400	Number of newspapers falls steadily
Periodicals (*issues*)	460	450	425	400	375	Number of periodicals falls steadily
Television (*programmes*)	33	33	55	65	75	Focus on satellite news but increase in hours per channel as storage rates fall
Radio (*programmes*)	4	50	100	200	200	Major expansion as national radio archive plans materialise
Websites (*sites*)	500	500	600	600	700	Steady rise in capture to include social media sites/feeds
Other media	0	0	50	50	100	Rise in areas such as citizen journalism
Connected (*issues*)	10	20	100	200	400	Rise in connecting to content from electronic services, partners and broadcasters
Totals	2,724	2,733	2,950	3,025	3,250	

Historical (i.e. the entirety of the collection)

Medium	2013	2014	2015	2016	2017
Newspapers (*issues*)	75,000,000	75,105,700	75,205,100	75,299,300	75,384,800
Periodicals (*issues*)	25,000,000	25,005,400	25,010,500	25,015,300	25,019,800
TV (*programmes*)	25,000	40,000	60,000	75,000	100,000
Radio (*programmes*)	10,000	30,000	50,000	75,000	100,000
Websites (*sites*)	0	150,000	300,000	700,000	1,400,000
Other media	0	0	2,500	5,000	10,000
Connected (*various*)	1,500,000	1,750,000	2,000,000	2,500,000	3,000,000
Totals	101,535,000	102,081,100	102,628,100	103,669,600	105,014,600

The aim in terms of content management is to reach a steady state for historical material by 2017, where the growth in the size of the newspaper collection is offset by the transition from print to digital. The aim for current material is to increase the number and range of sources to be made available to researchers, through a managed balance between collecting and connecting, emphasising diversity of forms and immediacy of access.

The categories below list the different news content types we offer, stating for each the current collection status, our collecting aim, and the action points.

1. Newspapers and periodicals

Status: The Library currently hold some 57,000 separate newspaper, journal, and periodical titles, of which 34,000 are newspapers, 23,000 journals and periodicals. The very approximate number of individual issues is 100M and the number of pages 750M. At present about 1,934 UK and Irish newspaper and weekly / fortnightly periodical titles are received per year, and 242 overseas titles. Access is provided onsite to an additional (and very approximate) 1.5M newspapers and news-related journals (individual issues) via subscription services, of which the main suppliers are ProQuest, Gale Cengage, Newsbank Readex and ProQuest. We point to at least 15M newspaper issues on freely-available sites listed on our website under Electronic Newspapers.

Aim: To continue to collect all UK and Irish newspapers under legal deposit, with a managed transition from print to digital collecting, but with the default position remaining print.

Action:

- Working with national publishers, use key national newspaper titles as a model to establish methods for print to digital transition under legal deposit while still maintaining 'heritage' collections of print newspapers
- Reduce print collecting of heavily advertising-based regional papers over next 3-5 years
- Begin collecting digital versions once the model for key nationals is established
- Maintain arguments for pre-print PDFs to be considered for legal deposit in any future legislation
- Consider collection of pre-print PDFs to give an aggregated online service to contemporary print content outside the reading rooms
- Collecting of periodicals and journals under legal deposit should follow same managed transition from print to digital
- No de-accessioning of the newspaper collection should take place, except potentially where there is duplication of print copies or extracts from print copies
- Maintain small, representative collection of newspapers from around the world, with default position being digital
- Review newspaper subscriptions budget with aim of separating UK news and news from other regions, responsibility for which should lie with respective curatorial areas

2. Television

Status: The Library currently holds 25,000 television news broadcasts from channels receivable free-to-air in the UK, recorded off-air since May 2010 under copyright exception, with 40 hours added per day from 15 channels.

Aim: To record and deliver access to representative content from all television news channels available free-to-air in UK from 2010 onwards, while connecting to historical television news archives.

Action:

- A dedicated service recording and archiving television news in the UK to be maintained, with emphasis on satellite news. There will be representative coverage from all freely-available channels in the UK, recording a minimum of 40 hours per day
- Explore the feasibility of a primary television access service via British Film Institute, complemented by BBC archive access, to include news programming, subject to a fuller assessment of the benefits, costs, resource implications and risks
- Explore with the BFI ways to open up access to its daily television news archives (going back to 1980s)
- Pursue connections with archive television resources, such as BBC Journal of Record and the Vanderbilt University Television News Archive
- Explore options for direct feeds from news broadcasters and news footage agencies
- Explore option for extending capture of television programmes to current affairs programming, news-related documentaries, discussion programmes, satire etc.
- Include broadcast access requirements in the specifications for the 'Universal Viewer'
- Explore with Ofcom and regional film archives the possibility of establishing network for capturing regional television news, including new regional television news consortia

3. Radio

Status: The Library has around 10,000 British radio news programmes, mostly recorded off-air in 1980s/90s or since May 2010 (BBC Radio 4 and World Service news). In September 2013 it will extend capture of radio news to 30 hours per day from 12 channels via the Broadcast News service. The Library is the official researcher access point for BBC audio archives, which includes news programmes.

Aim: To capture through off-air recording a substantial proportion of the UK's radio news output as part of an emerging national radio archive offering, while continuing to preserve and make accessible heritage radio collections.

Action:

- Seek radio industry and governmental approval for comprehensive archiving of Ofcom-licensed radio stations (c.600), financed through industry support

- Capture radio news extensively as part of national radio service utilising radio industry technology (Radio Monitor) and/or Broadcast News technology.
- Initiate pilot projects focussing on community radio, demonstrating value of associated research tools (particularly speech-to-text), targeting audiences in social sciences
- Improve BBC audio archive access by expanding upon the current Pilot Service
- Explore with BBC options for digitising the Library's BBC radio news scripts collection and linking this data to BBC /programmes records

4. Web and social media

Status: With non-print legal deposit legislation, introduced in April 2013, now in place, the Library aims to capture 4.5M .uk websites through an annual domain crawl. Additionally the Library plans more intensive, frequent curation of websites in three areas: selected events (four or five per year), key sites in curatorial subject areas (c.250) and news-based sites (c.500). A policy for capture of social media does not exist as yet, with technical challenges constraining ambitions in the short term.

Aim: To capture selected news-based sites crawled on a high-frequency basis, as well as an annual UK web crawl, including multimedia content as far as possible; to capture selected examples of news-based social media

Action:

- Establish specialist collection of few hundred web news titles for frequent domain crawl (ideally on a daily basis), with selection overseen by Legal Deposit Web Archiving Prioritisation Group in consultation with news curation team and other legal deposit libraries. Such sites will include newspaper titles, blogs, local media initiatives, news organisations etc.
- Use news-based websites as focal point for increased capture of online moving image and sound content, within non-print legal deposit regulations (although moving image and sound are excluded from the Act, the 2013 Regulations allow for the capture of audio-visual material as a feature within the main body of a work rather than as its main purpose)
- Test capture of news-based social media (particularly Twitter) by capturing feeds of journalists (professional and citizen), news organisations, plus ad hoc events via hashtags, special web pages etc.
- Track developments in online news production, particularly mobile news, with a view to extending capture under legal deposit
- Investigate potential for a news media gateway or portal within the context of the web redevelopment project

5. Other media

Status: The Library holds many other media which can contribute to a wider sense of what is news, including photographs, diaries, oral history recordings, maps, posters, letters and other manuscripts. It recently licensed the ITN image archive, now discoverable via Images Online. There are news media held in other collections which the Library can cross-link to, such as newsreels and cartoons. There are also types of news media not collected by any UK institution where the Library can build up strength, notably the area of citizen journalism video production.

Aim: To collect and connect to a range of content beyond the traditional understanding of what constitutes news, testing the viability of such an extension of service through pilot projects.

Action:

- Build up a collection of citizen journalism videos, through donation by individuals and developing selection and preservation arrangements with Web Archiving
- Extend range of external electronic news services to which the Library connects to cover newsreels (e.g. British Pathe), cartoons (e.g. University of Kent's Cartoon Centre)
- Open discussions with the BBC for special access to BBC Monitoring, its world news monitoring service (http://www.monitor.bbc.co.uk), on a service basis or more ambitiously on an extended access basis
- Include business databases such as EBSCO and Factiva in onsite electronic news offering.
- Build on partnership with British Universities Film & Video Council to provide enhanced access to its television, radio and newsreel databases
- Create an online directory of world news sources

6. Digitisation

Our aim is to extend access through digitisation. Web, television and most radio news content will be acquired in digital form, as increasingly will be the case for current newspapers and periodicals under the planned transition from print to digital for legal deposit intake. For historic newspapers and periodicals, digitisation is essential to deliver primary access to onsite and remote users. Our policy is to focus digitisation efforts on out-of-copyright newspapers, with emphasis on the mostly highly-used and academically-important parts of the collection. This is to be achieved through a mixture of public service and commercial activity, notably the arrangement with brightsolid to digitise 40M newspaper pages (approximately 5% of the collection) over a ten-year period, with free access onsite and access through subscription for remote users.

The brightsolid arrangement is proving to be of huge value to the Library and its users. Free in the British Library reading rooms and in the national library reading rooms of Wales and Scotland the www.britishnewspaperarchive.co.uk can be accessed via online micropayments and subscriptions. The content has transformed information gathering for genealogists who are accessing seven million pages of fully searchable content from the early eighteenth century through to 1950.

Through a partnership between brightsolid and Cengage Gale tranches of content are no available to institutions worldwide

While primary access to newspapers will be through digital and microfilm surrogates, it is important to recognise the importance for some researchers of engagement with original formats.

Digitisation activities for 2013-2017 will include:

- Continuing arrangement with brightsolid, with 8M further pages expected to be digitised by March 2015, achieving critical for UK researchers, particularly in field of family history
- Investigating new partnerships and funding opportunities for areas of the collections unlikely to be selected by brightsolid (but which could nevertheless provide the platform for their access), such as trade/specialist periodicals, tied into identifiable researcher needs
- Supporting the NEWSPLAN programme in its plans to digitise its microfilm set of regional newspapers, possibly using brightsolid / British Newspaper Archive as outlet
- With 16 partners, and over a 3-year period, contributing to target of 10M newspaper pages to be discoverable via the Europeana portal as part of the Europeana Newspapers project
- Extracting radio news broadcasts from off-air recordings made in 1980s/90s as part of digital preservation of sound collections
- Collaborate with other partners to digitise other non British newspaper content from the Newspaper collection

7. Discovery and access

A combined news media strategy presupposes combined discovery and access. The current state of things is a long way off from that ideal. The Library's Explore service is the planned discovery mechanism, but the different media have been catalogued at different levels, and there is no immediate way to search or browse the news media on any particular day, except for the individual British Newspaper Archive and Broadcast News services. The newspaper catalogue itself is a tool for finding bound volumes, boxes or microfilm reels, not for linking up individual issues (or stories).

News medium	Individual search	Combined search	Access
Newspapers	Newspaper Library search on Explore, British Newspaper Archive + third party services	Explore: Title level	Onsite + remote (British Newspaper Archive only)
Periodicals	Newspaper Library search on Explore, British Newspaper Archive + third party services	Explore: Title level	Onsite + remote (British Newspaper Archive only)
Television	Broadcast News	Explore & SAMI: Item level	Onsite only
Radio	Broadcast News	Explore & SAMI: Item level	Onsite only
Web	UK Web Archive	Explore: Item level	Onsite only
Other media	Images Online	None	Onsite + remote

Developing an integrated news discovery service will be a huge challenge, and the work required will need to be measured against the advantages it will bring to researchers. News requirements will need to inform future policy for discovery, but central to requirements will be item-level consistency across the different news media and the ability to search by day or other time period. The ideas below suggest ways in which to improve discovery in the short to medium term.

- Employ a 'news' tag to identify Library items as 'news' for easy discovery in Explore
- Develop a news search option on Explore, with 'news' tag, date searching and links to viewable/playable digital media
- Invest in speech-to-text technologies to open up access to radio and television news (where subtitles are not available)
- Evaluate and pilot methods to improve discovery and navigation such as OCR of background text, scene segmentation, scene characterisation, story segmentation (TV and radio news)
- Pursue licences which will enable to offer access to news media remotely and on readers' devices in the Reading Rooms, with protection and authentication systems to support this
- Encourage app development (through initiatives such as BL Labs) to create tools for the integration, discovery and reuse of news media sources
- Develop a British Library news data model, on Linked Open Data principles, which can be used to identify digital objects as 'news' and to build up linkages with partner organisations
- Develop a demonstrator which shows the research value of integrating diverse news media, built around a current news theme and a significant historical news event or subject

- Review how metadata policies and resource discovery systems might facilitate a re-imagining of how the British Library could present its collection in a time/date-based manner, facilitating a more comprehensive concept of ‘news’