ASSESSING NEWS DATABASES

Image from the *South China Morning Post* (September 30, 1948) available in Pro-Quest Historical Newspapers..

www.crl.edu

In This Issue

CRL's longstanding commitment to building and preserving newspaper collections continues. In the last five years alone, CRL and the Area Materials Projects working under the CRL umbrella collected, digitized, and micro-filmed newspapers from over 30 countries.

To multiply the impact of these efforts, however, CRL has now begun to gather and provide to libraries information to support local decision-making on investment in developing and maintaining library newspaper holdings and in the purchase of commercially available news databases. CRL's ICON database (page 2) is now the largest source of information on print and microform holdings at major newspaper repositories, and on the scope and completeness of content in "trusted" digital databases like CRL's *World Newspaper Archive* and the Library of Congress's *Chronicling America*.

In "Toward Greater Transparency in News Databases" (page 6), Maria Smith outlines how CRL will assess the sustainability and integrity of major newspaper databases and digital libraries. This assessment will enable libraries to make more informed decisions about print holdings, and invest with greater confidence in news databases. James Simon and Diane Ryan's analysis of academic news databases (page 9) illuminates the extent to which news broadcast transcripts are covered in major databases like *Factiva* and *LexisNexis Academic*.

Together these efforts are designed to provide libraries actionable information about the infrastructure and organizations we rely upon to ensure our researchers long-term access to important news content. This will be essential in a time when library control of news content is becoming more critical than its custody.

—Bernard F. Reilly, Jr.
President

The ICON Database: New Information for Decision-making on Newspaper Digitization and Preservation

Bernard F. Reilly, Jr.
President, Center for Research Libraries

Image of ICON database display for the *Times of India*, showing issues held and gaps in the aggregate holdings represented in ICON.

The Challenge

Effectively managing and providing access to historical newspapers are matters of consequence for libraries. Academic libraries and many national libraries invest considerable sums to digitize newspapers to make them more accessible to historians and other researchers. And each year, research libraries in the aggregate spend millions of dollars to purchase databases of digitized historical newspapers from commercial publishers. The pressure to provide these resources is rising: News content is in high demand from researchers, who are now able to use sophisticated software and applications to mine the large bodies of data-rich text that newspapers represent.

At the same time, original back files of newspapers are increasingly threatened. National libraries and many major academic libraries, long able to maintain extensive and bulky collections of these files, are now under intense pressure to repurpose scarce storage space to provide public amenities such as exhibits, study rooms, and social spaces. In other instances, newspaper back files are imperiled by inherent vice: they were printed on poor-quality paper and often stored under unfavorable environmental conditions. As a result, libraries must make consequential decisions daily about whether to preserve, conserve, reformat, and even retain these types of important materials.

Unfortunately, few precise figures are available to inform these decisions. However, The Andrew W. Mellon Foundation recently awarded the Center for Research Libraries major funding to expand CRL's collecting and analysis of data on archived and digitized newspapers and, based on that data and analysis, to promote coordinated, strategic action by libraries and consortia.

CRL's Role in Newspaper Preservation and Acquisitions

CRL today provides data and analysis to enable academic and independent research libraries to make informed decisions about their own acquisition and management of news materials locally. CRL's white papers on the lifecycle of electronic news (2011), the adoption of the web by African newspaper publishers (2013), and the comparative coverage of news broadcast transcripts by the major aggregators (2013) are a few of the resources designed to help illuminate the complex landscape of digital news for librarians.

CRL now also brokers the terms of purchase and licensing agreements between many of its libraries and the vendors and publishers of databases and electronic resources, through which most scholars today obtain access to primary source
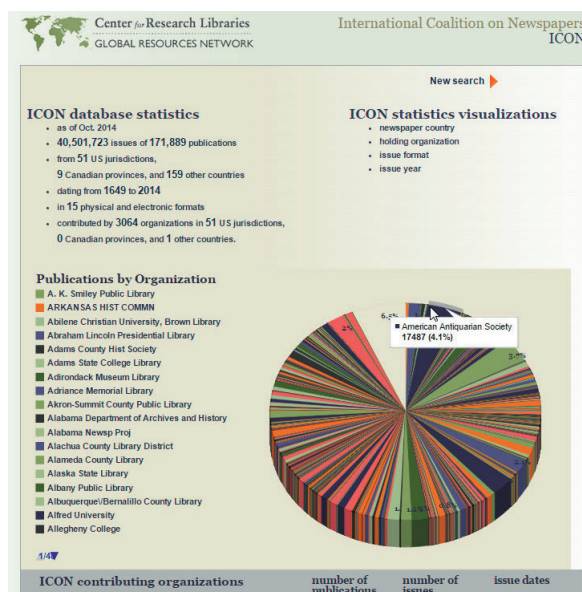
materials. CRL publishes critical reviews of major newspaper databases in CRL's online *eDesiderata* platform.

Newspaper Collection Data and Analysis: the ICON Database

One of the major resources CRL provides for librarians is the *ICON Database*, produced under the auspices of the International Coalition on Newspapers program. ICON is a registry of information on the hard copy, microform, and digitized holdings of US and foreign newspapers held by several major US research libraries. The largest single repository of information about such holdings, ICON currently contains records for over 172,000 newspaper titles, published in 51 US jurisdictions, 9 Canadian provinces, and 159 other nations. The publication dates of these holdings range from 1649 to 2014. At present, ICON lists over 40 million issues of these titles represented in the print and microform newspaper holdings of CRL and several major US research libraries, the extensive print newspaper holdings of the American Antiquarian Society, and digitized newspapers in two databases considered by CRL to be trustworthy and persistent: LC's *Chronicling America* database and the Readex *World Newspaper Archive*.

For each newspaper title, the ICON database displays:

- **Publishing history of the title**, on an issue-by-issue basis, which is generated using an algorithm developed by CRL from sampling, extrapolation, and hands-on data generated as a by-product of the digitization process, microfilming process, and/or intensive shelf-reading.

- **Names and characteristics of repositories/databases holding the title**, and any or all of the formats in which the title is held.

- **Holdings of the accredited repositories** in microform and paper down to the issue level, obtained through CRL harvesting or publisher direct submission. CRL harvests this information from the Chronicling America website using an API, and obtains direct submission from Readex and the American Antiquarian Society.

- **The contents of the major "trustworthy" databases** on an issue-by-issue level, obtained through direct submission from the publisher and through CRL harvesting.

We envision three practical uses of ICON data, namely to support the following decision scenarios.

1. Strategic digitization

For libraries and publishers investing heavily in digitizing newspaper content, there is currently no single source of information on newspapers that have already been digitized. Reliable, granular information on the location of complete print and microform holdings, which might serve as potential source materials for digitization, is also scarce. A comparison of the contents of the *Chronicling America* database and the vast collection of early US newspapers held by the American Antiquarian Society, using ICON, reveals what a small portion of the latter have been digitized.

2. Collection management

ICON data can confirm the existence of original and reformatted copies, indicating not only the completeness of a library's holdings but also the conditions under which those holdings are maintained. Such information can be relevant to library decisions on whether to retain locally held original copies of a given title, or whether to invest scarce resources in their conservation, attempt to fill gaps, or implement better security, environmental conditions, or controls on handling.

3. Investment in database purchases and acquisitions:

Gaps in coverage are a common flaw in databases of digitized newspapers. ICON's increasingly reliable inventory of the editions of many titles actually published, in the form of an issue-by-issue publishing history, provides a frame of reference for judging the completeness of a given newspaper database.

Historically, this kind of information has not been available. Bibliographic and holdings information for newspapers, where available, tends to be incomplete, general, or unreliable. Holdings reported to utilities like OCLC are often described in summary terms, without noting gaps. Even this information can sometimes be out of date, largely because of the labor required in obtaining such data, not to mention the scarcity of information on the publishing histories of newspaper titles.

However, the widespread mass digitization of newspaper collections is now creating new opportunities for capturing this data: the digitization process creates issue-by-issue, article-by-article, and even page-by-page data. If we capture that data, as CRL is doing, it becomes possible to describe newspaper titles with unprecedented detail and granularity.

Funding from The Andrew W. Mellon Foundation is now enabling CRL to enhance the ICON database to handle data at a more granular level, and to expand our ability to collect data on newspaper titles held and digitized by other major world libraries and by key commercial news database publishers.

During 2014–15, CRL is creating the necessary protocols and an ingest pathway for automating the importing of issue-level information into ICON. These will enable new digitization projects to contribute to the ICON database data in various common metadata schemas and packages. CRL is developing software to parse and normalize issue-level data on newspaper holdings in hard copy, microform, and digital formats, and to enrich publication histories based on existing ICON data.

Challenges in Securing the Metadata

CRL faces real challenges in obtaining the detailed metadata necessary to fuel ICON analysis. While the Library of Congress exposes *Chronicling America*

metadata to open harvesting using an API, we hope to convince other major US and European libraries to follow suit. The capability to create such metadata is already present in the widely used CCS docWorks software, and in other newspaper digitization applications. We hope that exposure of this data will become the norm in library digitization.

The cooperation of other electronic publishers will also be critical to the usefulness of ICON, and that cooperation is not yet assured. Several electronic publishers have already agreed to submit or expose for ongoing CRL harvesting the issue-level bibliographic and descriptive metadata for titles in their existing databases and digitization pipelines. Readex provides CRL metadata on the *World Newspaper Archive* databases as a condition of its cooperative agreement with CRL. As of October 2014, ProQuest has tested delivery of metadata for one long-running title (*Times of India*), and is considering additional contributions. Gale and East View have agreed to explore submission of data but we have not yet seen results.

CRL maintains that exposure of metadata at the issue level should be considered by libraries a basic prerequisite of transparency and thus trustworthiness in commercial databases. CRL will prevail upon the other major national site-licensing consortia, such as JISC (UK), the Deutsche Forschungsgemeinschaft (DFG, Germany), and CRKN (Canada) to help us convince publishers that making their metadata harvestable by CRL or submitting it to ICON is in their best interest.

Toward a Coordinated, Rational Strategy

With support from The Andrew W. Mellon Foundation, CRL will also work with several major US, UK, Canadian, and European organizations to make future newspaper preservation and digitization more systematic and strategic. In early 2015, using ICON data, CRL will conduct a comparative analysis of the coverage of world newspapers by the major digitization efforts to date. That analysis will examine and evaluate the major "trustworthy" databases, and identify significant weaknesses, gaps, and areas of overlap and duplication. The analysis will identify by title and country of origin many newspapers not yet digitized, and of intrinsically high risk (e.g., titles neither digitized, nor micro-formatted, nor widely held; titles historically prone to vandalism or theft; titles published during eras of highly acidic paper, etc.).

The findings of the analysis will then be shared with representatives of the major actors in newspaper digitization: the Library of Congress and the National Endowment for The Humanities, JISC, the Europeana Newspapers partnership, the DFG, interested national libraries, and the major database publishers. The findings will be the basis for deliberations at an international "summit" on newspaper archiving and digitization that CRL will convene in conjunction with the meeting of the IFLA News Media Section, in April 2015. There, representatives of national and academic libraries, consortia, electronic publishers, and others will weigh the findings of the ICON analysis and their implications for further mass digitization of newspapers.

The summit will also be an opportunity to decide on acceptable and achievable norms and protocols for sharing data about newspaper digitization projects; perhaps even create the outlines of a sustainable and mutually advantageous "division of labor" between the commercial publishers, national libraries, and major library consortia on the future digitization of international newspapers.

A common agenda for newspaper digitization and preservation that is rational, strategic, and achievable would create much-needed clarity around the mass-digitization undertaken by libraries and publishers. It would minimize duplication of library and publisher investment, and ultimately optimize the usefulness of news databases for scholars. ❖

Toward Greater "Transparency" in News Databases: Assessing Policy and Practice of Digitized Newspaper Repositories

Maria Smith
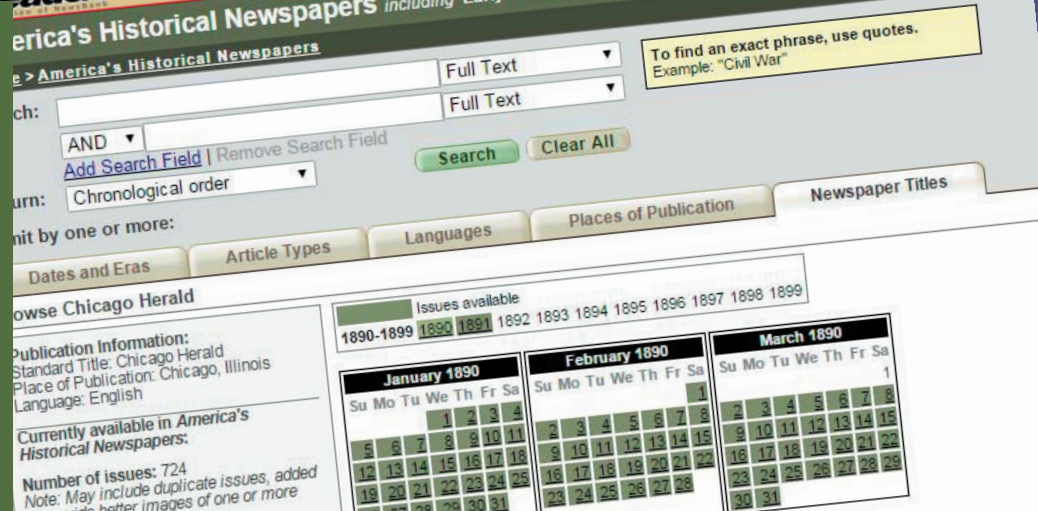*News Database Analyst*
Center for Research Libraries

Image from Readex's *America's Historical Newspapers* collection showing issues held for the *Chicago Herald.*

The expansion of access to digitized newspapers affords many opportunities for researchers and libraries today. Since 2001, a growing and diversified body of providers—national and academic libraries, state and public institutions, publishers, third-party commercial producers—have brought scalable digital technologies to bear on the provision of legacy newspaper collections.

However, as with the challenges of exposing structural metadata on available collections (*see previous article*), the diversity of players and the multitude of platforms and preservation systems employed in the process have provided little transparency for libraries seeking information on the sustainability and likely persistence of digitized news content. To meet this challenge, with the support of The Andrew W. Mellon Foundation, CRL is gathering detailed, comparative information on the infrastructure, content management, and digital repository platforms maintained by commercial publishers and libraries that support digitized newspaper collections.

CRL's Assessment Framework

CRL has created an assessment framework to evaluate the major platforms and repositories, as well as the practices and capabilities of the supporting organizations with respect to their ability to provide long-term access to digital news resources. The two primary purposes of the framework and the assessment efforts are:

1. to inform library decisions about the purchase and licensing of newspaper databases, and
2. to support strategic library and publisher investment in the development, implementation, and management of digitized newspapers programs.

CRL has a rich background in facilitating informed decision-making at academic and independent research libraries. Through its reviews of electronic resources, CRL provides objective information on major databases and digital collections of high interest for licensing by participating CRL libraries. CRL also creates provider profiles on its *eDesiderata* site, which analyze the business practices and sustainability of the resource providers. At the most rigorous level, CRL conducts in-depth assessments of digital archives that preserve research materials of interest to the CRL community. CRL reports are independent, critical analyses based upon information obtained from the repositories themselves, from reliable independent sources, and in some instances through formal audits.

These initiatives have informed CRL's development of a new framework for assessing digitized newspaper repositories. The assessment is based on two major sets of

criteria for trusted digital repositories: *Trusted Repositories Audit and Certification: A Checklist*[1] and ISO 16363 : Audit and certification of trustworthy digital repositories[2]. However, the framework under development for digitized news repositories seeks to strip down the checklist criteria to a level of assessment that is credible and actionable, while employing a rating scheme less rigorous than full certification. The framework will provide "actionable intelligence" about repositories and programs, presented in the form of structured repository profiles that will be widely shared with libraries and other newspaper digitization stakeholders.

In assessing the major libraries and publishers conducting newspaper digitization, CRL seeks to answer the following questions:

1. Does the organization provide adequate support, staffing, and financial resources to perpetuate the long-term sustainability of the repository?
2. Does the organization demonstrate sufficient plans and policies for the selection, digitization, and storage of content?
3. Does the organization follow content-management practices that utilize a workflow model demonstrating valid ingest and archiving processes as well as facilitating sufficient intellectual control of content?
4. Does the organization utilize systems that adequately protect repository content and enable the monitoring of content integrity?

To fully answer these questions, CRL is documenting the policies, platforms, and technologies used by the libraries and publishers to manage digitized historic newspaper content. The information CRL is gathering falls into the following broad areas:

ORGANIZATION APPARATUS

To understand how well digitized newspaper content is managed by a given publisher or organization, CRL will examine the overall mission, project scope and documentation, general structure of the organization, and contingencies in place for ensuring long-term content integrity and persistence.

PROGRAM PLANNING AND MANAGEMENT

To identify the key requirements for program implementation, CRL will examine the planning and preparation that goes into managing the organization's repository. CRL will gather documentation of formal decision-making responsibilities, guidelines for content retention and disposition, rights and permissions regimes, specifications for digitization and enhancement of content, and the organizational framework or model for storage and access to repository content.

CONTENT MANAGEMENT

To assess the effectiveness of the workflow in promoting sustainability, CRL will survey the technical processes and workflows involved in submitting/acquiring, ingesting, and maintaining content in the repository. CRL will also assess the nature and quality of repository metadata and how it is maintained; the ability to express and expose structural metadata for purposes of assessment and discovery; and the ability to structure hierarchical relationships between content data objects and content sources (i.e., documenting the provenance of content).

TECHNICAL COMPONENTS

CRL will analyze the systems and technologies utilized to support the repository and sustain a digitized newspaper program.

1. *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* February 2007. Version 1.0. OCLC and CRL http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

2. *International Standard ISO 16363: 2012 Space data and information transfer systems – Audit and certification of trustworthy digital repositories.* Geneva, Switzerland: International Organization for Standardization.

The Framework in Action

CRL has begun applying the framework in the profiling and rating of several library programs and commercial newspaper products, including: the Library of Congress National Digital Newspaper Program, the National Library of Australia Newspaper Digitisation Program, and Readex's historical newspaper databases. CRL consults open-source information (staff papers, publications, conference presentations, annual reports, budgets, agreements and contracts, website content, repository samples, and media coverage), confers with administrative and technical managers from the participating organizations, and employs the help of subject matter experts within the CRL community of libraries to build accurate and current assessment profiles.

The rating scheme proposed by CRL is based on the level of transparency of the repository operations, program accountability, and trustworthiness as demonstrated by the organization. In rating the programs, CRL will provide an objective measurement of the ability of a provider to demonstrate credible policies, practices, and infrastructure in support of the program, through such metrics as:

- Organizational policies that explicitly define the financial and bureaucratic commitment to the newspaper program;
- Plans for preserving the institution's digital/digitized content over time;
- Presence of appropriately trained staff thoroughly engaged in the newspaper program;
- Evidence of content selection priorities, offering rationale and guidelines for carrying out selection procedures with awareness to accountability;
- Documented workflows and metadata specifications;
- The content and metadata retrievable through digital platforms and repository samples indicating the level of compliance with standardized content management practices;
- Specifications for open source and/or proprietary software and hardware employed by the repository; and
- Evidence and details on the application of backup and security tools used in the repository's storage and dissemination systems.

The assessments and profiles will identify the credible aggregators of historical newspapers, illuminate community standards and best practices, and analyze the risks posed to the persistent access to digitized newspapers over time. These tools will develop the knowledge base for CRL's efforts to build broad consensus among libraries, publishers, and commercial providers for the further digitization of newspapers globally. ❖

Broadcast News Transcripts in Academic News Databases

Diane Ryan

Former Project Coordinator, World Newspaper Archive

James Simon

Vice President, Collections and Services

Center for Research Libraries

Image from *Infotrac Newsstand* showing a broadcast transcript for *PBS Newshour* (November 22, 2012).

Despite the growing diversity of news sources, television remains the most popular source for Americans to follow news.[1] Interest in television news broadcast content for research purposes also continues to grow as tools to access and interpret broadcast become more widely available. Transcripts of selected news broadcasts have been accessible in print and microform since the mid-20th century. The *Vanderbilt Television News Archive* (*VTNA*) began recording television news programs in 1968 for academic research. New sources of information from television news—such as the *UCLA Library Broadcast NewsScape* and the *Internet Archive TV News Archive*—continue to transform the research potential of this critical media.

In 2013, the Center for Research Libraries conducted a comparison of six academic databases regarding the extent and accessibility of broadcast transcripts for national and international news. The survey aimed to assess the coverage and scope of broadcast media (television, radio) in the major research resources subscribed by CRL member institutions, measuring the relative strengths of each product in terms of depth of coverage and uniqueness. CRL also sought to assess the overlap in content among the major providers, and to compare findings to those of community-driven efforts such as the *Vanderbilt Television News Archive* and the emergent *UCLA Library Broadcast NewsScape*.

The six databases assessed during the 2013 study are listed below[2]:

- *Access World News* (NewsBank)
- *Factiva*
- *InfoTrac Newsstand* (Gale Cengage)
- *LexisNexis Academic*
- *Newspaper Source* (EBSCO)
- *ProQuest Newsstand*

CRL's full assessment revealed a number of unique characteristics for each product. In addition to content assessment, CRL evaluated the ease of use and the ability to quickly identify and access information in broadcast transcripts using both simple and advanced search functions.

A summation of the content coverage is reported below.

News Coverage Compared

Factiva and *LexisNexis Academic* are large databases with several thousand research resources, most of which are non-news sources. *Factiva* focuses on
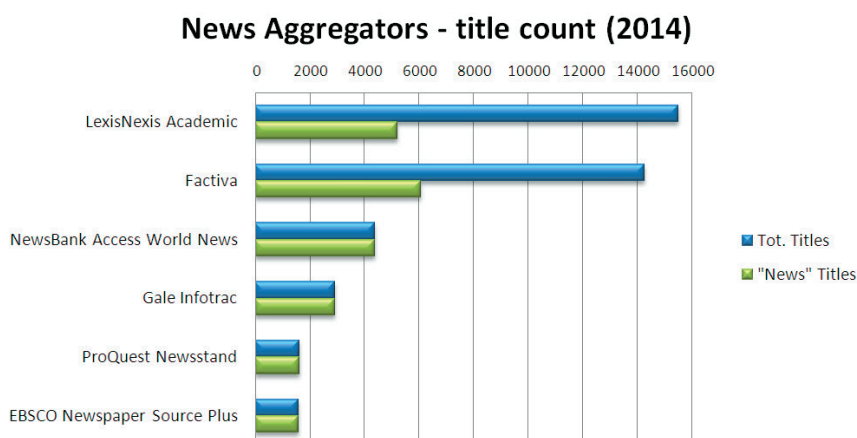
1. http://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/ and http://www.journalism.org/2013/10/11/how-americans-get-tv-news-at-home/

2. The statistics for the various products were collected between May and June 2013. As such, they represent a snapshot of the content coverage of the products, and may not represent current figures in 2014.

resources that influence business. It provides access to a reported 35,000 sources, including more than 3,300 US, regional, and international newspapers; approximately 1,200 newswires; news websites; media transcripts; blogs; and multimedia. *LexisNexis* provides access to legal resources in addition to major news sources. It reports more than 15,000 total sources, of which an estimated 3,000 are classed as newspapers and an additional 1,000 sources from newswires, web-based news publications, and news transcripts.

The other four products focus primarily on news: *Access World News* has nearly 4,400 news sources (3,200+ newspaper titles); *InfoTrac Newsstand* has 2,900 titles (1,600 of which are newspapers); *ProQuest Newsstand* and EBSCO's *Newspaper Source Plus* both contain approximately 1,600 news sources, of which nearly 1,200 are full-text newspapers.

These figures, compiled in 2014 from the most recent product source lists, show continued growth in news titles across all aggregators compared to 2013 figures. This suggests an ongoing commitment by the providers to expand the availability of news sources for academic research.

News Aggregators - title count (2014)

Broadcast Transcript Availability

All six products provide access to news broadcast transcripts, though the depth of coverage and time span varies. Using vendor source lists (2013), the following numbers of sources for transcripts were identified:

LexisNexis Academic	462
Factiva	430
Access World News (NewsBank)	229
Newspaper Source (EBSCO)	136
InfoTrac Newsstand (Gale Cengage)	102
ProQuest Newsstand	24

The numbers above do not necessarily indicate relative total size of collections, as each product lists sources in a different way. For instance, *Access World News* lists mostly broadcast stations included, whereas others list individual programs. Providers were not consistent in how programs and sources were listed.

For all products, transcripts are predominantly in English, gathered primarily from US sources. Generally, all products present "full-text" coverage of the aggregated broadcasts, with only a handful of sources represented as selected coverage or abstracts.

Assessment of Broadcast Source Content

Factiva reports offering transcripts for 430 news programs, spanning approximately 1990 to present. *Factiva* provides television and radio transcripts from the major national and cable news networks, including ABC, CBS, CNN, FOX, NBC (with CNBC and MSNBC), NPR, and PBS. *Factiva* works with providers such as ASC Services (formerly Morningside Partners) in the provision of content, as well as direct supply from publishers.

Factiva includes international broadcast content from the Australian Broadcasting Corporation, BBC Monitoring, CNN International, CTV Television Network (Canada), Deutsche Welle (Germany's international broadcaster), and Euronews (France). *Factiva* appears to be aggressively adding international and foreign-language broadcast content, such as from BFM TV and Radio Monte Carlo (France); multiple Arabic-language sources such as *al Qarra TV* (France, focusing on news of the African continent), *al Arabiya* (UAE, Gulf Region), Middle East Broadcasting Center (MBC) in Saudi Arabia, and *CNBC Arabiya*; and *TV Express Search* service covering multiple stations in Japan.

Transcript coverage is extensive, with national commercial network programs extending back to the 1990s (NBC coverage begins November 1989, CBS from 1990, and ABC from 1994), and cable news offerings dating back to 1997. Of the 430 programs/sources listed, approximately 40 percent are currently updated.

Factiva contains numerous programs not listed in other products' source lists. However, the variance in how aggregators choose to list source titles makes difficult to ascertain true uniqueness among products. *Factiva* lists multiple program titles—mostly non-current—from CNN and ABC News Now (ABC's 24-hour news channel) not listed by the other aggregators. Most unique content offerings, however, are primarily from the international sources. *Factiva* is the only product listing Deutsche Welle content, as well as the Arabic and French sources described above. In 2013, it added coverage from several Russian-language broadcasts from Ukraine.

LexisNexis Academic (LN) reportedly includes broadcast transcripts from more than 450 programs, primarily from mainstream US national news sources, but also from Canada, Australia, UK, France and the Asia-Pacific region. Like *Factiva*, *LN* sources some content directly from publishers, but more frequently through third-party transcription services including the Federal News Service, CQ-Roll Call, and ASC/Morningside. International coverage is provided through sources such as BBC Monitoring and Euronews, and recently *LN* has begun to incorporate international sources from other non-European areas, including South Africa (Summit TV, serving the South African business community) and India(*Economic Times Now*).

Coverage of broadcast transcripts is extensive, with many sources extending back to the 1990s or earlier, depending on the service or publisher (ABC coverage for *World News Tonight* in *LN* extends back to 1980, CBS programs back to 1990, CNN to 1992, NBC to 1993). Of the 462 programs listed, approximately half are reported to be updated currently.

Content unique to *LexisNexis* appears to be primarily in the international news arena, including *Channel NewsAsia*, the aforementioned *Summit TV*, and selected programming from the CTV Television Network in Canada. Video content is provided by third-party sources including TVEyes, a media-monitoring service including full-text transcripts, and NewsLook, a video platform that aggregates news from multiple sources.

Access World News (AWN) lists nearly 230 transcript sources, though its source list includes only the name of the broadcasting station rather than the individual

program titles the other vendors include as sources. *AWN* contains content from national providers including Bloomberg, CBS, CNBC, CNN, FOX News, MSNBC, NBC, NPR, and PBS. In addition, *AWN*'s source list includes approximately 200 local broadcast stations from around the United States, incorporating local news coverage by affiliate stations of the major networks, independent stations, and others.

AWN also features international content, including several unique sources: BBC selected transcripts, Independent Television Network (Sri Lanka), OrissaTV.com (India), and RBC TV (Russia).

Content coverage of national broadcasts begins roughly around 2003, with the addition of local station coverage beginning in late 2006. NPR content extends back to 1990.
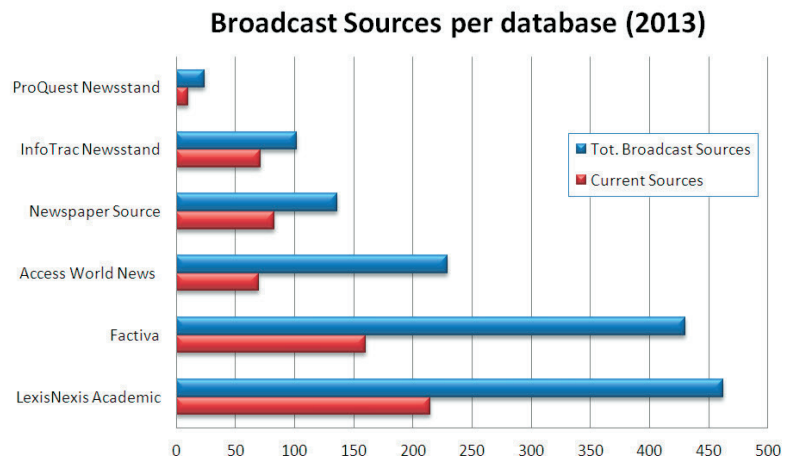
Of the 230 listed sources, 160 (or 70 percent) are reported as "not current." Nearly all of these sources are local broadcast stations (the exceptions being CNBC and CNNfn, which have limited coverage in the database, and CNN en Español, which was indexed from 2003 to 2011).

InfoTrac Newsstand includes approximately 100 transcript sources, provided primarily through CQ-Roll Call, Inc. Content includes programs from CBS, CNN, FOX, and Bloomberg. *InfoTrac* also aggregates selected transcripts from NBC, NPR, PBS, and the Federal News Service. *InfoTrac* includes AP transcripts beginning from January 2000.

Of the 102 sources listed, 71 titles are reported as currently updated. However, most of the content uniquely held by *InfoTrac* is no longer currently updated, such as local content from McClatchy-Tribune Information Services. The product contains selected sources for audio and video, such as *Wall Street Journal This Morning* and *National Review Online* content. *InfoTrac* lists Critical Mention, Inc. and ShadowTV as sources for local broadcast and Bloomberg video content, though details are not available.

Newspaper Source lists 136 different transcript sources, 60 percent of which are still currently supplied. Coverage includes programs from ABC, Bloomberg, CBS, CNN, FOX, MSNBC, NPR, and PBS. *Newspaper Source* also includes content from the Australian Broadcasting Corporation and Canada's CBC Television. NPR content dates back to 1998, with other broadcast stations picking up between 2000 and 2003. CNN content begins around 2005.

ProQuest Newsstand presently only includes transcripts from BBC Monitoring, a division of the British Broadcasting Corporation that monitors and reports on mass media worldwide. Coverage dates back to 2003 for most BBC content.

Broadcast Sources per database (2013)

Overlap in Broadcast Coverage

In comparing the various title lists, CRL found 729 unique listings among the six products. Approximately 40 percent of the titles were held by two or more vendors. The amount of overlap identified is as follows:

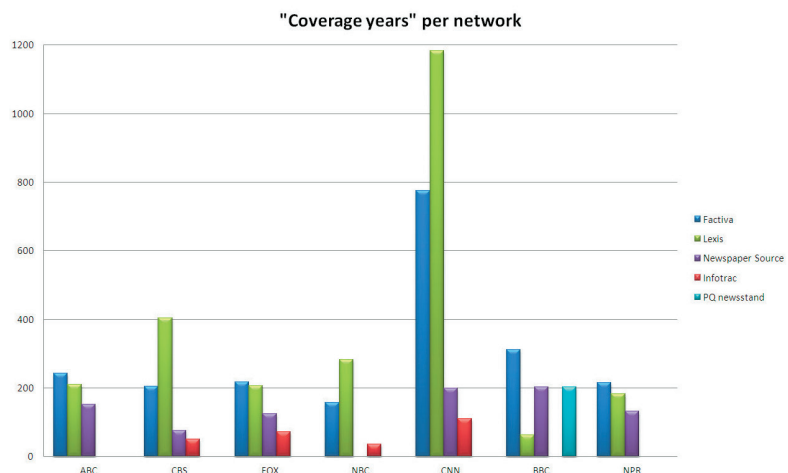| Number of titles held in common	
Four or more vendors	26
Three vendors	74
Two vendors	180
One vendor	453

As suggested above, comparing coverage among the databases is difficult due to inconsistencies in how each aggregator reports its source list. Titles are not listed consistently, nor with any authority control. Not all sortable fields match from product-to-product. Most significantly, some vendors list only the broadcast publisher as only one source, while others list each program from the publishers as individual source titles. Thus, in all likelihood the percentage of overlap is higher, but without deeper content analysis the reviewers were unable to conclusively align the title lists.

Depth of Coverage Compared

To compare depth of coverage, CRL sampled 42 titles held by three or more vendors to determine which vendors maintained the deepest collection of broadcast articles. CRL found that for all titles sampled, *LexisNexis* and *Factiva* consistently had the longest runs of coverage.

Though coverage varied widely from title to title, generally *LexisNexis* was found to have deeper backfiles of content than *Factiva*, often having older runs of transcripts, ranging from a few additional months up to 14 years (the average difference was four years). Of the 42 titles sampled, *LN* had deeper backfiles for 20 of the titles. For eight titles, *Factiva* had deeper content, though generally extended by only one or a few additional years. Fourteen of the titles had more-or-less equal coverage in the two databases.

Comparing by major broadcaster and extent of runs, *Factiva* and *LexisNexis* are competitive in terms of content offering. Looking solely at total years of coverage across all programs of each major network, *LexisNexis* is stronger in coverage for CBS, NBC, and CNN, while *Factiva* holds a slight lead in coverage for ABC, FOX, NPR. These strengths vary, however, from year to year. *LN* is considerably stronger for early coverage of transcripts, while *Factiva* ramped up its collecting efforts in the 2000s. Both products maintain robust aggregation of all major networks from 2010 to present.

Content Comparison: Television News Archives

CRL examined coverage of *Vanderbilt Television News Archive* and the emergent *UCLA Library Broadcast NewsScape* to compare their news coverage to the text-based databases above. While these databases focus more on audiovisual capture and presentation than on broadcast transcripts, they are significant and growing efforts that provide an alternate means of accessing broadcast news programming.

While *VTNA* does not provide the same type of source list as the commercial aggregators, coverage may be summarized as follows:[3]

Regular News Programs

- *ABC Nightline:* August 5, 1968–present
- *CBS Evening News:* August 5, 1968–present
- *NBC Evening News:* August 5, 1968–present
- *CNN: WorldView:* October 2, 1995–November 3, 2000
- *Wolf Blitzer Reports:* February 1–December 31, 2001
- *NewsNight:* November 5, 2001–Oct 28, 2005
- *Anderson Cooper 360:* November 1, 2005–present
- *FoxNews Reports:* January 15, 2004–present
- *ABC Nightline:* March 24, 1980–September 12, 1988: occasional coverage
- *ABC Nightline:* September 12, 1988–present: comprehensive coverage.

Special Reports

The term "Special Reports" refers to news coverage of significant events broadcast outside the scope of the regular evening news programs. This part of the collection focuses on US presidential politics, including political conventions, election coverage, and speeches and press conferences of the president currently in office. The Special Reports collection also includes coverage of major national and world events and major military conflicts involving the US. In addition to the networks described above, *VTNA* contains special reports from the following broadcasters:

- PBS
- MSNBC
- CSPAN
- CNBC
- Univision
- Bloomberg

VTNA does not cover local news programming, or "news magazine" programs such as *60 Minutes* and *20/20.*

UCLA's *NewsScape* contains a variety of national news programs and local news shows from the Los Angeles area. It began its coverage in January 2005, covering approximately a dozen news programs from ABC, CBS, FOX, NBC, and KCAL (an independent station featuring local broadcast news). It expanded its programming coverage in October 2006 to include CNN, Fox News, MSNBC, and additional national and local programs. It added Spanish-language content in August 2007. *NewsScape* relies on closed-captioning texts for its full-text searching, rather than transcripts of abstracts.

3. http://tvnews.vanderbilt.edu/coverage.pl

While *NewsScape* coverage does not go back as far as *VTNA*, its coverage of national news programming appears to be more extensive than Vanderbilt, including daytime programming and major cable news networks. *NewsScape* also includes extensive coverage of news-related programs of an entertainment nature (*The View*, *Entertainment Tonight*, *EXTRA*, *The Colbert Report*, *Saturday Night Live*, and so on). From a content perspective, *NewsScape*'s coverage of local broadcast, independent stations, and news-related programming sets much of its content apart from the other databases. ✥

Center for Research Libraries Staff Contacts
(800) 621-6044

President
Bernard F. Reilly x 334
breilly@crl.edu

Administrative Services Specialist
Yvonne Jefferson x 319
yjefferso@crl.edu

Member Liaison and Outreach
Services Director
Mary Wilke x 351
mwilke@crl.edu

Vice President, Collections
and Services
James Simon x 324
jsimon@crl.edu

Director of Technical Services
Amy Wood x 327
awood@crl.edu

Director of Information Systems
Patricia Xia x 341
pxia@crl.edu

Head, Access Services
Kevin Wilks x 314
kwilks@crl.edu

Head, Stack Management
Bethany Bates x 339
bbates@crl.edu

Head of Communications and
Development
Virginia Kerr x 265
vkerr@crl.edu

Communications Specialist
Gwen Ihnat x 289
gihnat@crl.edu

Global Resources Program Contacts
(800) 621-6044

Director
James Simon x 324
jsimon@crl.edu

Global Resources Network and AMPs
Program Manager
Judy Alspach x 323
jalspach@crl.edu

News Database Analyst
Maria Smith x 322
msmith@crl.edu