Submitted on: 11/08/2014

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

Bernard F. Reilly, Jr.

Center for Research Libraries, Chicago, USA
reilly@crl.edu

Abstract:

Libraries invest heavily to digitize to newspapers and spend millions to purchase databases of digitized historical newspapers from commercial publishers. At the same time, they are required to make weighty decisions about the disposition of original newspaper back files, which are rapidly disappearing. Unfortunately, these decisions are based on little hard information.

The Andrew W. Mellon Foundation recently awarded the Center for Research Libraries (CRL) major funding to expand its collecting and analysis of data on archived and digitized newspapers and, based on that data and analysis, to help create a common international agenda for newspaper digitization and preservation that is rational, strategic and achievable. At the center of this effort is CRL's recently expanded ICON database.

Keywords: Newspapers, preservation, digitization, digital humanities, metadata.

The Challenge

Managing and providing access to historical newspapers are matters of important consequence for libraries. Academic libraries and many of our national libraries invest considerable sums to digitize newspapers to provide access to historians and other researchers. Each year, research libraries in the aggregate spend millions to purchase databases of digitized historical newspapers from commercial publishers. The pressure to do this 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 newspaper back files are rapidly disappearing. Many national libraries and 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, little hard information is available to inform these library purchase and retention decisions. However, the Andrew W. Mellon Foundation recently awarded the Center for Research Libraries (CRL) major funding to expand its 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

The Center for Research Libraries was established in 1949 as a cooperative venture of ten US Midwestern research universities, to acquire and preserve critical or at-risk source materials for researchers. Today a consortium of over 200 U.S., Canadian, European, and Asian libraries, CRL is now the largest existing cooperative preservation and collection development enterprise, and focuses its resources on collecting and preserving primary historical and cultural evidence produced in both digital and print formats.

CRL today also 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 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's white papers on the lifecycle of electronic news, the adoption of the Web by African newspaper publishers, and the comparative coverage of news broadcast transcripts by the major aggregators, and critical reviews of major newspaper databases in CRL's online eDesiderata platform, are some of the resources designed to help illuminate the complex landscape of digital news for librarians.[1]

Newspaper Collection Data and Analysis: the ICON Database

One of the major CRL resources for librarians is the ICON Database, produced under the auspices of the International Coalition on Newspapers program at CRL.[2] ICON is a registry of information on the hard copy, microform and digitized holdings of U.S. and foreign newspapers held by several major U.S. research libraries. The largest single repository of information about such holdings, ICON currently contains records for over 172,977 newspaper titles, published in 51 US jurisdictions, 9 Canadian provinces, and 159 other nations. The publication dates of these holdings range from 1649 to 2012.

[1] See the eDesiderata platform at: http://edesiderata.crl.edu/
[2] ICON Database is at: http://icon.crl.edu/

At present, ICON lists over 39 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:

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

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

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

d. Issues of the title held in 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:

1. Strategic digitization decisions: Many libraries and publishers are investing heavily in digitizing newspaper content, to serve the high demand for such source materials from researchers. Yet there is currently no single source of information on newspapers that have already been digitized. Reliable, granular information on where complete print and microform holdings reside, holdings which might serve as potential source materials for digitization, is scarce. Using ICON, a comparison of the contents of the Chronicling America database and the vast corpus early U.S. newspapers held by the American Antiquarian Society, reveals what a small portion of the latter have been digitized.

2. Decision-making on collection management: ICON data can be used to 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 chronic flaw in databases of digitized newspapers. ICON's increasingly reliable inventory of the issues of many titles actually published, in the form of a day-by-day, week-by-week, year-by-year publishing history, provides a frame of reference for

judging the completeness of a given newspaper database. Scope and completeness is a qualitative feature of a given database, and will be reported as such in CRL's eDesiderata database reviews.

Historically, this kind of information has not been available. Bibliographic and holdings information for newspapers, where available in the major utilities, tends to be incomplete, general, or unreliable. Holdings reported to utilities like OCLC are often described in terms of summary holdings, with gaps undescribed, and even this information can sometimes be out of date. This is largely because of the labor required in obtaining such data, and also owes something to 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 the production and capture of this data: the digitization process creates issue-by-issue, page-by-page, and even article-by-article data. If we capture that data, as CRL is doing, it becomes possible to describe newspaper titles with unprecedented detail and granularity.

The Mellon Foundation funding is 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.

In 2014-15 CRL will create the necessary protocols and an ingest pathway for automating the import of issue-level information into ICON. The pathway and protocols will enable new digitization projects to contribute to the ICON database data in various common metadata schemas and packages. CRL is currently 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.

CRL will also develop new scripts to standardize and enhance holdings data from other accredited databases and sources. This will enable CRL to increase the granularity of ICON information about reformatted papers, and permit a more rigorous comparison of the newspaper databases with collections that have not yet been reformatted, identifying gaps and materials at risk.

Challenges in Securing the Metadata

There are challenges, however, that CRL us facing in obtaining the detailed metadata necessary to fuel ICON analysis. The Library of Congress exposes Chronicling American metadata to open harvesting using an API. We hope to convince the major U.S. and European libraries to follow suit, and make issue-level metadata harvestable by CRL. 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 commercial publisher Readex provides CRL metadata on the World Newspaper Archive databases as a condition of its cooperative agreement with CRL. But the cooperation of other electronic publishers will also be critical to the usefulness of ICON, and that cooperation is not yet assured. CRL is now attempting to secure agreement from several electronic

publishers to submit or expose for ongoing CRL harvesting, issue-level bibliographic and descriptive metadata for titles in their existing databases and digitization pipelines. Gale and East View have agreed to submit that data but we have not yet seen results. As of June 30, 2014, discussions with ProQuest were inconclusive.

CRL maintains, however, 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 (U.K.), the 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 Greater "Transparency" in News Databases

CRL will also enlist publishers to provide CRL relevant information about the systems and platforms they use for managing and exposing their digital content. This is to enable CRL to assess the suitability of those systems to provide long-term access to digitized news. CRL's reviews of major news databases, such as ProQuest's Times of India and Latin American Newsstand, available in the CRL eDesiderata platform, appraise databases in terms of the scope, quality, and integrity of their content, and the extent to which they accommodate current scholarly practice.[3]

CRL has now begun to produce provider profiles, which offer an independent perspective on publisher business models, infrastructure, resources, and practices. This year CRL will undertake risk analyses on several major publishers of news databases, in an effort to determine how credible their management of digital content is. This activity is an outgrowth of CRL audits of digital preservation repositories, which have to date provided risk assessments of Portico, HathiTrust, CLOCKSS, and Scholars Portal, based on the TRAC checklist.[4]

Using a new rating scheme for publishers, CRL will evaluate the platforms and digital content management capabilities practices as they reflect on the publisher's ability to maintain digital resources for the long term. In August 2014, CRL will share the preliminary ICON scheme with members of the IFLA Newspaper Section, and will solicit comment on the rating system from a wider set of stakeholders. That review will produce specifications for necessary changes to the CRL rating scheme, and perhaps additional metrics.

However, it is clear already that the ability and willingness of a publisher to produce and provide metadata at the issue level – indicating the presence of gaps, if any -- on the contents of their databases will be a factor in CRL's assessment of their control over their digitized content, and therefore a measure of the trustworthiness of their platform.

[3] The Times of India and Latin American Newsstand reviews are posted in eDesiderata at: http://edesiderata.crl.edu/resources/times-india and: http://edesiderata.crl.edu/resources/latin-american-newsstand

[4] For CRL certification reports, see: http://www.crl.edu/archiving-preservation/digital-archives/digital-archive-reports. The Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), is the principle tool used by CRL in its auditing and certification of digital repositories. For TRAC, see: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac

Toward a Coordinated, International Strategy

With support from the Andrew W. Mellon Foundation, CRL will also work with several major U.S., UK, Canadian, and European organizations to make future newspaper preservation and digitization more systematic and strategic. In 2015, using ICON data, CRL will perform 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, 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 Deutsche Forschungsgemeinshaft, 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, we hope in cooperation with IFLA's newspaper section, in 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; and perhaps for creating 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; minimize duplication of library and publisher investment; and ultimately optimize the usefulness of news databases for scholars.

Looking Ahead

The effort I have described is a work in progress. A tremendous amount of development has already been made possible with support from CRL's U.S. and Canadian member libraries, the U.S. Institute of Museum and Library Services, National Endowment for the Humanities, and the Andrew W. Mellon Foundation. However, there is much to be done and further success will ultimately depend upon the cooperation of librarians and publishers. There is reason for optimism, and for confidence that the time has come for international, concerted action, resting on a solid foundation of reliable data.

Acknowledgments

The data and analysis for this presentation were provided by Amy Wood, Director of Technical Services at the Center for Research Libraries, and the project director for development of the ICON database.