Center for Research Libraries
GLOBAL RESOURCES NETWORK

Framing a Common Agenda for Newspaper Digitization and Preservation: an ICON Summit
Report and Outcomes

On April 14th, 2015 the Center for Research Libraries (CRL) convened an international gathering at the National Library of Sweden, "Framing a Common Agenda for Newspaper Digitization and Preservation: An ICON Summit". Participants at the Summit included approximately 30 members of the newspaper digitization community: representatives from national and academic libraries, commercial producers and aggregators of news content, and representatives of CRL, Jisc Collections and LIBER. The purpose of the gathering was to exchange views on the present state of international newspaper digitization and preservation, and to identify ways to better align future library and commercial efforts, investment and funding in these areas.

This report is a summary of the proceedings and discussions at the event, and the next steps CRL has identified as a result of the Summit.

Background

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 from commercial publishers databases of digitized historical newspapers. At the same time, many libraries are under growing pressure to trim collections and reduce storage of bulky and increasingly embrittled collections of newsprint. As a result, libraries must make consequential decisions daily about whether to preserve, conserve, reformat, and even retain these types of important materials.

In 2014 the Center for Research Libraries received funding from The Andrew W. Mellon Foundation 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, publishers and consortia. CRL approached this task in two ways: a) by putting in to place a framework for identifying "trustworthy" digital platforms and repositories for digitized newspapers managed by libraries and publishers, and b) by assessing the current scope and coverage of historical world newspapers in those repositories. The goal of the evaluation framework and data assessment is to provide greater transparency for libraries seeking information on the comprehensiveness and long-term integrity and accessibility of digitized news content.

CRL held the ICON Summit in conjunction with the IFLA International News Media Conference in Stockholm, Sweden. Two studies were disseminated prior to the conference:

- A Comparative Analysis of Newspaper Digitization to Date
- Digitized Newspaper Repository Assessment Scheme (proposed)

These documents laid the groundwork for the presentations and discussions.

Session 1: "The State of the Art:" A Comparative Analysis of Newspaper Digitization to Date

The first session of the Summit was devoted to CRL's analysis of the scope of newspaper digitization efforts to date, using granular data to measure the extent to which those efforts have covered the history and geographic scope of newspaper publishing in the major world areas. **James Simon** (Vice President, Collections & Services, CRL) presented the findings of the report: "The "State of the Art": A Comparative Analysis of Newspaper Digitization to Date." For the study, CRL used title-level information available from major newspaper digitization projects, and issue-level information from other newspaper projects and databases, aggregated by CRL in the ICON Database of International Newspapers (http://icon.crl.edu). The ICON database contains issue-level newspaper holdings data for digitized newspapers, as well as information on titles held in print and/or microform.

Among the many findings reported were the following:

- *Newspapers digitized to date fall mostly within a relatively narrow time frame in the history of news publishing*. While newspaper publishing came of age during the late eighteenth and early nineteenth century, the predominant focus of most digitization efforts is the late nineteenth and early 20th century. In fact, within that period coverage from efforts surveyed peaks between 1890 and 1918. Content availability rapidly declines after 1923 (the U.S. copyright "cliff"), although some European libraries have scanned newspapers dating as late as approximately 1945. Overall, the overwhelming majority (87%) of materials digitized thus far dates from prior to the mid-twentieth century
- *Most newspapers digitized to date were reformatted from existing microfilm, rather than original paper copies*. Chronicling America and the World Newspaper Archive content is almost entirely from microfilm source materials.
- *Overwhelmingly, the newspapers digitized to date were published in the United States, the U.K., and Western and Northern Europe.* From a geographic perspective, national and local priorities in these regions are the primary driver of digitization activities. This has resulted in a relative dearth of coverage of news resources from world areas with limited infrastructure or limited capacity to digitize and disseminate their own heritage.
- *In publicly and commercially funded digitization, English-language content dominates, and there is little access to digitized news content published in the non-Western, less commonly taught languages*. Unfortunately, this occurs at a time when newspaper publishing in regions like South Asia and Sub-Saharan Africa is growing rapidly.

This uneven landscape of available content is due to multiple factors. Factors include copyright restrictions; national collection mandates of publicly funded projects; local priorities and financial interests.

Simon demonstrated the difference between title-level analysis, which provides information on the presence of titles in any given year, and issue-level assessment, which can depict the "saturation," or depth of coverage, of titles in a given year. Using the *Chronicling America* data harvested by CRL into the ICON database, Simon showed how issue-level data can present a more nuanced picture of coverage

over time. In this case, coverage of papers for the time period 1888 through 1918 appeared more balanced than the title-level assessment initially suggested.

Even at the scale of holdings in ICON (over 42 million issues from nearly 170,000 newspaper titles) the database as yet contains only partial coverage of the world's news output. CRL is urging commercial and noncommercial providers to contribute issue-level data. Using ICON data contributed by Readex for the World Newspaper Archive, Simon demonstrated how regional and geographic coverage of digitized newspapers changes over time. ICON data can also be used to compare digitized collections to holdings in print and microfilm to assess how much has been digitized relative to the amount of all materials held by trusted repositories.

Because only a small fraction of the world's news content has been made electronically accessible, CRL sought to determine which titles remain most "at risk" and why. Categories of content identified as most endangered include:

- *Newspapers existing only in print format.* Digitization efforts, emphasizing providing access over preservation per se, are making accessible the least endangered materials, while many print collections—especially the highly-acidic wood-pulp from the 20th century—remain largely unpreserved and at risk due to use, deterioration, or neglect.
- *News material from areas that do not have strong preservation infrastructure or capacity.* There is little market incentive for commercial publishers to digitize materials from developing world regions. Moreover, in certain areas, there may be the added threat of past or current regimes or other actors actively trying to suppress or destroy published documentation.
- *Materials still under copyright.* Due to library and publisher aversion to risk, relatively few contemporary newspapers are digitized. In order to clear rights for digitizing copyright-era newspapers these require the substantive cooperation of news publishers, limiting the ability of libraries and cultural heritage organization to make headway in this area.

Limitations of the Existing Data

CRL's analysis is based on the limited amount of data available to CRL to date. Efforts to harvest information on digitized news content are intended to inform libraries on what content has been preserved, reformatted and/or digitized, where it is held, and how likely it is to remain available or future researchers. CRL also endeavors to identify materials most at risk or deserving of preservation attention. Yet, CRL faces real challenges in obtaining the detailed metadata necessary to fuel its ICON analysis. While CRL maintains that exposure of issue-level metadata and information on the source or provenance of newspapers digitized by libraries and publishers should be considered a key indicator of the trustworthiness of a given newspaper database, such data is not currently available from many library and commercial databases.

Simon noted that access to this kind of data is particularly important today because of the growing challenges in preserving newspaper materials for the long term. One challenge is the decline of public support/funding at the national and state levels; pressing economic imperatives are drawing funding away from memory organizations' digitization and microfilming activities. Moreover, the most at-risk

materials are not commercially attractive to publishers and are therefore less likely to be targets for reformatting by those organizations. It was agreed that public-private partnerships will be required to achieve comprehensive digitization of the world's newspapers.

Session 1: Discussion

Digitization Costs

Digitization project costs vary based on institutional priorities: for instance, the Library of Congress's Chronicling America cost per page is very high, but their priority was to establish a digital archival effort, mandating high standards for metadata, archival processes and quality control. Another major hidden cost of digitization is the cost of building and maintaining within the organization expertise on the history of news publishing and knowledge of the published oeuvre. It was pointed out that this is an asset in which both commercial vendors and libraries are required to invest.

The cost of digitization often does not take into account the ongoing costs of maintaining the digitally reformatted content, such as hosting, storage, and format migration. CRL asked if these costs are being accounted for when planning for digitization. One participant representing a European national library suggested that their digitization funding includes projected costs for an initial three years of project management.

Management of Content Rights

One participant remarked, "At the end of the day, it's all about rights"; to understand digitization decisions and strategies, it is necessary to know the legal constraints within which such strategizing takes place. Participants discussed the wide variation in copyright and access conditions within which different countries and organizations work. In reality, database vendors and national libraries often have only limited rights to make digitized news content accessible. Rights granted database vendors by newspaper publishers and suppliers of content are often of limited duration and/or accompanied by restrictions on uses such as text mining. Similarly, the rights granted national libraries under legal deposit arrangements, governing the extent to which they can provide access to content harvested or otherwise obtained in digital form, vary from one country to the next.

In the national libraries of Europe, conditions for legal deposit of newspapers seem to cross the spectrum: many libraries still require that publishers deposit print editions (such as at the British Library) or grayscale microfilm for preservation. However, some libraries have begun accepting PDFs or have implemented systems for e-legal Deposit, such as the case in Finland. In terms of providing access on an international level, libraries explained that often content may be made available only at the national level. A summary of European legal deposit conditions for web content by country can be found at http://www.netpreserve.org/legal-deposit. One participant shared the sentiment that in Europe and worldwide, more international cooperation should be encouraged to establish guidance for rights management in library digitization.

CRL posed the question to the commercial representatives regarding their ability to share more information on the rights to content that they possess. From the commercial perspective, this would be

exposing competitive information; while CRL suggested that libraries making content decisions on digitization or preservation would be better informed if such information on rights were available.

Commercial representatives suggested it would be difficult to describe conditions for all newspaper content maintained by a provider. There are complications involved in decoding and describing contracted rights to newspaper content; contracts may precede the digital age and, therefore, do not address digitization/digital access. Moreover, there is not always clarity in early contracts when it comes to rights and what a publisher can do; a provider may have rights to a digital asset but could be missing necessary rights for certain access capabilities such as text and data mining. Uncovering and clarifying conditions of contracts is often done on an as-needed basis.

Overall, participants agreed that greater disclosure of information about the rights held by vendors and national libraries and the limitations on same would be useful, but that the real complexities needed to be addressed.

Prioritizing digitization and preservation of "at-risk" materials

CRL posed the question to the participants: Does this discussion change any of your priorities for digitization and preservation? Are the libraries present concerned, for example, about the loss of non-domestic materials such as those from the developing regions? What are the possibilities for adopting less traditional funding models in order to digitize and preserve these and other at risk materials?

Most representatives from national libraries felt that their digitization efforts were likely to follow established paths. The group did discuss public-private partnerships and funding for expanding digitization of materials from non-Western, developing worlds. There are acknowledged limits to what a commercial provider can do, but more funding and investment from public institutions might create possibilities. Most participants agreed that digitization purely for preservation's sake would be a hard sell: in order to execute such projects there would need to be some market demand. Digitization priorities should be backed up "by researcher's voices," and assessing commercial viability of a collection and content needs can be a lengthy process.

As an alternative solution to national funding, the group discussed the possibility of bringing together interested academic community members to subsidize the costs of digitizing marginalized material. While there may not be prospects for financial support through the "taxpaying community," there are academic communities that will provide support for specialized collections, as evidenced by CRL's area studies groups: http://www.crl.edu/collaborations/global-resources-programs.

End Session 1

Session 2: Measuring the Effectiveness and Value of Digitization Programs and Repositories

The second session of the Summit introduced the topic of measuring the effectiveness of digital newspaper programs through a formal assessment scheme for evaluating digitized newspaper repositories.

Bernard Reilly introduced the session by discussing the rationale behind the Digitized Newspaper Repository Assessment Scheme, a set of metrics CRL is developing to assess or measure the trustworthiness of major databases. The principal motivation for developing such a scheme has to do with the growing reliance of scholars on digital resources, as opposed to print and microform collections, and growing pressures on research libraries to contain or reduce their collections storage costs. The resulting divestment by libraries of newspaper collections has made digital access often the only form in which some newspapers are available to researchers. While information is being created at a faster rate than ever before, it is also being lost at a faster rate. In order to know that a library's digital collections are functional we need to know more about how they are supported.

A second motivating condition for developing an assessment scheme is the growing role that commercial providers play in providing long-term access to historical newspaper content. Because public funding for newspaper digitization is becoming scarcer, libraries will increase their reliance on third parties to collect, host, and make accessible scholarly materials. Under these circumstances, libraries must know more about—and have confidence in—a publisher's ability to manage content for the long term. In the past, corporate mergers and vendor bankruptcies have led to the disappearance of newspaper microform. Additionally, unclear rights held by an organization pose a threat to future use, as do temporary or limited rights to publish and maintain the content. CRL believes that a healthy assessment regime could provide transparency and create confidence in the major digital libraries, both commercial and public.

The rating framework proposed by CRL for digitized newspaper repositories measures how reliable a repository will likely be in the long term. CRL's long history in auditing and certification of digital preservation repositories on behalf of its members provided the basis and methodology for the assessment scheme presented today. The metrics are based on accepted community standards, such as the Trustworthy Repository Audits and Certification (TRAC) and ISO Standard 16363: Audit and Certification of trustworthy digital repositories. However, the newspaper repository scheme proposes a less rigorous process than full certification, designed instead to facilitate a "light audit" of a digital repository and to provide the basis for publishing information about the repositories in the form of an online "profile."

Maria Smith (Digital Repository Analyst, CRL) presented the specifics of the proposed Digitized Newspaper Repository Assessment Scheme. Aside from the TRAC checklist and ISO 16363, the assessment scheme draws upon well-established standards for digital repositories and newspaper digitization: Deutsche Forschungsgemeinschaft's P*ractical Guidelines on Digitisation,* Library of Congress'

T*echnical Standards for Digital Conversion of Text and Graphical Materials, and* Educopia Institute's Chronicles for Preservation *Guidelines for Digital Newspaper Collection Preservation Readiness*.

Smith walked participants through the primary and secondary criteria to be used in measuring and rating repositories. As a case study for the discussion, CRL produced a Digitized Newspaper Repository - Sample Assessment to suggest the types of documentation used in the assessments and to demonstrate how ratings may be applied in practice.

Session 2 Discussion:

In general, participants expressed interest in the proposed scheme and broadly agreed that it was a useful development. Fewer participants were certain that they would be able to undertake a full assessment based on what they heard at the session, but were willing to consider it. Some of the questions and themes that emerged from the discussion are presented below.

Scope

Would the scheme be applied to microform and print repositories, or only to digital repositories? At present CRL proposes to assess digital repositories only under this initiative; we are not currently looking at how microfilm holdings or print holdings are maintained, as that would require entirely different types of audit criteria and methodology.

Uses

Who are the intended users of the assessment reports? The assessment scheme was developed as a means of providing information for libraries in making collection management and reformatting decisions, and in deciding on the value and integrity of commercially available databases. The profiles will inform libraries on the longevity of database content, influencing decisions on investment as well as preservation and digitization of their holdings.

Evidence

Participants wondered about the data that supports and facilitates this kind of assessment; CRL maintains a list of documentation and sources of evidence that map to the various criteria points used to evaluate the level of compliance with the assessment principles, which was not presented for the Summit given the extensiveness of the material. A practical example of documentation used in a TRAC audit is the document checklist built by Scholar's portal during their certification process, available at: https://spotdocs.scholarsportal.info/display/OAIS/Document+Checklist

Certification limitations

Participants were curious about how accurately a certification reflects the level of risk to content in a repository. As CRL knows from previous audits, certifying a digital repository has limitations, and being "certified" does not negate the existence of threats to content; within digital repository management, risks will continue to exist. We therefore seek to describe those risks. It is also important to note that

some issues or shortcomings may not result in non-certification; instead we will identify and notify stakeholders of those existing risks.

How can we evaluate financial wellbeing for the long term? CRL has previously reviewed 3 years of financial statements and projections to measure how the organization accomplished and met what they projected; this information can represent stability, longevity and changes to funding over time. Now, we may seek to review 5 years of financial statements and projections.

In considering long term viability of an organization, it is also important to review the mechanisms for decision making related to its ability to meet the needs of libraries. It is important in the assessments that we answer the questions: who are making the key decisions at the organizations in terms of the kinds of systems they build and the content they acquire?

As a representative of a commercial organization pointed out, contracts with publishers often establish that the newspaper publishers maintain ownership of content throughout its lifecycle, how does this factor in to the assessment when assessing rights management by the repository's information? CRL is concerned with the rights that the aggregator maintains that prove they are legally enabled to manage content over a defined time period, we are particularly concerned with rights as they related to and effect perpetual access conditions.

Intrusiveness

Participants discussed the risks involved in third-party assessments of major database and platforms. These included threats to system security in disclosing proprietary technical information and details about system architecture. In its audits CRL has used non-disclosure agreements with organizations, and a secure internal enterprise wiki platform for storing and managing sensitive information

There was further concern expressed regarding how CRL guarantees protection of sensitive competitive business intelligence to ensure it will not be leaked, especially when considering the risks posed by employee interaction with sensitive data. Information designated by repository managers as privileged or confidential is not disclosed by CRL without permission, as CRL is bound by NDAs to prevent information leakage. This is of the utmost importance to CRL, as any violation of the agreement would be incredibly costly and therefore detrimental to the organization.

Objectivity

Participants also discussed the possibility of subjectivity in how assessment metrics are applied, considering that repository practices might be described as "sufficient", "capable" or "consistent" arbitrarily. A related question arose about how assessment activity can be supported.

It was pointed out that CRL's accountability to its member community—the North American academic and independent research libraries and the scholars served by those institutions—provides its standing as an arbiter of trustworthiness in scholarly resources. As a result CRL undertakes audits of repositories of interest to its community, and bases its audit charges on the level of that interest. Moreover, fees

charged for audits cover only the direct costs, i.e., the human resources, travel, services, and materials needed to undertake the audit/assessment.

Benefits of Assessment

Discussion turned to how the various organizations represented at the Summit might benefit from the scheme. Some participants believed that the metrics could be used as a convincing argument for funding digitization, or technical changes and improvements in a digital library or repository. It was also pointed out that disclosure of granular metadata on the contents of a newspaper database could help publicize additions made and gaps filled in such databases. One participant thought that independent assessment of their digital library could be useful as a way to identify the risks and threats to their repository that are not currently apparent. This affirms CRL experience with audits of digital repositories like CLOCKSS and Portico, where the process of conducting the audit has often led to improvements in repository policies and processes, and a better sense of the kind of information libraries want disclosed.

End Session 2

Next Steps:

Aside from recent efforts by CRL, very little hard data is available on the repositories and digital libraries of digitized newspapers. Information about the gaps in library holdings of newspapers in print and microform, issue-level data on the contents of major newspaper databases, and qualitative information about the platforms on which major newspaper databases and digital libraries are maintained and accessed are all needed to enable libraries and publishers to make prudent, strategic decisions about digitization and preservation. CRL will work to obtain and expose more, granular metadata from the major libraries and vendors on their digitized newspaper holdings. CRL will also endeavor to enlist the cooperation of digital library program directors and publishers of news databases in disclosing information about the platforms, repository architectures, and relevant business practices on which their sourcing, management and dissemination of important digitized newspapers rely. Disclosure of such information in the ICON database and in CRL's repository profiles will promote greater transparency and more informed investment in the newspaper digitization and preservation.

In the last fifteen years libraries and publishers have digitized an enormous amount of newspaper content: many millions of pages are now available in open access digital libraries like Europeana Newspapers and Chronicling America, and in commercially-produced databases. This is a boon to scholars and researchers, enabling remote electronic access to important primary source materials on a 24/7 basis from anywhere in the world. This immense digitized corpus, however, represents only a small fraction of the number of newspaper titles that are preserved in libraries. Moreover, if one considers the digitization of newspapers a means of preservation, we have not succeeded in capturing a significant amount of the materials that are most at risk. Those endangered materials include secondary-market newspapers published in the second half of the twentieth century; newspapers from all eras that have

not yet been microfilmed; and newspapers published in developing or historically unstable world regions and newspapers published in non-Western languages. Therefore, in its digitization efforts, CRL will prioritize at-risk materials: primarily newspapers published after 1945 in developing or historically unstable world regions and newspapers published in non-Western languages. In this effort, CRL will identify and foster opportunities for private-public partnerships, and will engage its community of academic and independent research libraries.

Support for the ICON Summit and preliminary analysis was provided through the generous assistance of the Andrew W. Mellon Foundation as well as from CRL member libraries.