IFLA International Newspaper Conference 2010

# IFLA International Newspaper Conference 2010

# Digital Preservation and Access to News and Views

## CONFERENCE PAPERS


Editors
Ramesh C. Gaur, Frederick Zarndt
D. R. Gupta, Kavita Gaur

Indira Gandhi National Centre for the Arts, New Delhi
IFLA Newspaper Section

IFLA International Newspaper Conference 2010

# Digital Preservation and Access to News and Views

CONFERENCE PAPERS

(25th – 28th February, 2010)

**Editors**

Ramesh C. Gaur

Frederick Zarndt

D. R. Gupta

Kavita Gaur

**Indira Gandhi National Centre for the Arts, New Delhi**

**IFLA Newspaper Section**

**Published by:** ***Indira Gandhi National Centre for the Arts***
***IFLA Newspaper Section***
February, 2010
New Delhi

**Printed by:** Om Laser Printers Delhi- 110009

# Table of Contents

# Editors' note

We are pleased to present to you some of the papers of IFLA International Newspaper Conference 2010: Digital Preservation and Access to News and Views being organized jointly by Indira Gandhi National Centre for the Arts and IFLA Newspaper Section from 25th to 28th February, 2010. The volume includes a total of 29 papers both invited as well as contributed by more than 45 authors from 9 different countries. These papers cover various aspects such as physical preservation, digital preservation of both traditional as well as born digital, digitization, online newspapers, newspaper as digital resource and 24X7 digital access to newspapers. Also explored are various issues concerning digital archiving of newspapers and the reading habits of users in the digital era.

The papers in this volume are presented with a view to triggering further discussions during the Conference. Later this volume may also help the participants as a reference tool in implementing some of the ideas discussed during the Conference. This is the first time that such an International Conference is being organized in India. Therefore, we believe the newspaper- libraries and archives in India would be able to derive much value out of the contents of this volume.

We have tried to include almost all the papers being presented in the Conference. However, due to the paucity of time some of the papers have not been included. We understand that a complete volume will be published by IFLA Newspaper Section as proceedings of the Conference in due course of time and some of the papers which have not been included here may be part of the proceedings.

The deadline for receiving of full text papers was 7th February, 2010. However, many of the papers have been received till 17th February, 2010. So, we were having few days to edit these papers. Despite our best efforts some errors may have crept in, so due apologies to Authors for any such errors.

We are grateful to all the speakers at the Conference for providing valuable inputs for this volume. Our sincere thanks to the Printer of this volume for timely printing and special thanks to Mr. R.S.Kaushik for his help. The typing assistance received from Mrs. Kiran Kapoor and Mrs.

Sunita Arora IGNCA is duly acknowledged. Last but not the least we are grateful to the Members of the Publication Committee of the above Conference, namely, Ramesh C. Gaur, Frederick Zarndt, Par Nilsson, Edmund King and Hartmut Walravens for their contributions in screening and review of proposals received for the above Conference. Without their efforts this volume may not have been possible.

Editors

# From the IFLA Newspapers Section Chair

This year's IFLA Newspapers Section conference, on the theme "Digital Preservation and Access to News and Views", takes a progressive look at news and newspapers. In keeping with the digital focus of Stockholm's August 2009 conference on "The Present becomes the past: Harvesting, archiving, presenting today's digitally produced newspapers", the 2010 New Delhi conference also focuses as much on collection, preservation, and access of digital news and newspapers as it does on these same activities for traditional print newspapers.

The reason for this is obvious: News is now created and produced digitally, even for print newspapers. Almost without exception, mid-sized or larger newspaper publishers distribute news via the Internet concurrently with their print issues. And although news continues to be distributed in print, some newspaper publishers such as the Christian Science Monitor have stopped printing daily newspapers; other publishers have eliminated the print issue altogether, and still other publishers are experimenting with distribution via devices such as Amazon's Kindle.

The Newspapers Section has in past been 'concerned with all issues relating to newspapers in libraries and archives, including acquisition and collection development, intellectual and physical access, storage and handling, preservation of newspapers and their contents'. Now it faces new tasks such as collection (harvesting) and preservation of online news and born digital newspapers and providing access to it as well as historical print newspapers. These tasks are not trivial.

With the daily or weekly print newspapers, one knows what must be collected and preserved – it has a physical instantiation -- even if one does not always have the resources to do so. However content in born digital news and newspapers and especially in online news can be changed and updated so rapidly that it is not possible to collect or preserve it with traditional methods and technologies. Furthermore digital news is fragile: Bit rot corrupts files. Digital photos and digital content are only a hard disk crash or a virus infestation away from destruction.

This present era of transition from traditional newspapers to digitally produced news offers both opportunities and challenges. New thinking, new methodologies, and new technologies are needed to cope with the challenges of digital news production and distribution. But with the seminal efforts already made by cultural heritage organisations around the world, with the further help of the librarians, researchers, practitioners, and archivists whose papers are collected in this volume, and finally with the further efforts of you the reader and conference attendee, these necessary innovations shall be discovered.

Frederick Zarndt, Chair
IFLA Newspapers Section
Coronado CA USA
February 2010

# From the Conference Director

It is indeed an honour to extend a hearty welcome to all the participants, speakers and eminent guests at IFLA International Conference 2010, being jointly organized by Indira Gandhi National Centre for the Arts and IFLA Newspaper Section from 25th to 28th February, 2010. This is for the first time that such an International Conference is being organized in India. When I was contacted by Mr. Hartmut Walravens former Chair, IFLA Newspaper Section for organizing this Conference at IGNCA, I was not sure of accepting this proposal, as neither did I have any background in Newspaper Libraries nor does IGNCA have any programme for preservation of newspapers. Later, I accepted it because I found this an opportunity to create awareness about preservation of and access to newspaper-based information services in India.

Initial response to the Conference information circulated by me on various listings was reasonably encouraging. But the kind of response I received in terms of participation, papers and sponsorship is amazing. Let me tell you that major cost expenditure with regard to this Conference has been taken care of with the help of sponsorships from the industry including registration fee of the participants and speakers. The delegates to the Conference are from libraries of more than 15 countries such as Library of Congress, British Library, National Library of Australia, National Library Board, Singapore, National Library of France and Denmark. Coming back to our own country, along with the major reputed newspapers i.e. The Times of India, The Hindustan Times, The Hindu, Indian Express, DNA, Deccan Chronicle, several regional newspaper libraries such as those of Matrabhumi, Eenadu, Rajasthan Patrika are also participating in this Conference. We have also received registrations from some media persons in this conference. Apart from Newspaper Libraries and Archives, we have representation from National Library of India and some important Universities of India. I am sure with such a broad-based participation, the Conference will be able to achieve the objectives set out for this event of international significance.

This Conference has been organized with the efforts and cooperation of many individuals and institutions. It may not be possible for me to name everyone. However I would like to put on record the critical support received from the following individuals and institutions. Let me begin

with IFLA Newspaper Section particularly Mr. Frederick Zarndt for having faith in me for organizing of this Conference. I am grateful to all the authorities of IGNCA, Officers and staff for their constant support and encouragement in the organization of this Conference. I sincerely thank all sponsors for their financial support for this Conference. I would like to specially put on record the support rendered by Mr. Vishal Salgotra and Mr. Bharat Joshi of Planman Technologies in the organization of this Conference. I am grateful to all my colleagues from media libraries, namely, Mr. Dharam Vir, Mr. R. Venkata Kesavan, Mrs. Pratibha Kaushik, Ms. Vijaylakshmi, Mr. Pranav Priyadarshni, Ms. Anita Pujari, Mr. K. Rajindrababu for their support in making this Conference a real success. Excellent efforts by all officers and staff of Kala Nidhi Division of IGNCA are duly appreciated and acknowledged. I also appreciate the support from the Ministry of Culture, Ministry of Home Affairs and External Affairs in granting permissions and clearances for foreign delegates. I would also like to express my gratitude to all the participants, speakers, Chairpersons of various technical sessions and Rapporteurs for agreeing to contribute to this Conference.

For the last few months I have been working for 18 hours a day despite a great personal loss (of my father). I had to devote a large amount of time which actually belongs to my family. My heartfelt thanks to my wife, Mrs. Kavita Gaur and to my daughters Ritu Gaur and Kanika Gaur for allowing me to use their time for the organization of this Conference. Without their support all this could not have been possible. Last but not the least, I am grateful to all those whose names I may not have included here but their efforts were nonetheless there in the successful organization of this Conference.

Ramesh C. Gaur
Conference Director &
Head – Kala Nidhi Division, IGNCA
IFLA International Newspaper Conference, 2010
February, 2010

# IMPROVING ACCESS TO DIGITIZED HISTORICAL NEWSPAPERS WITH TEXT MINING COORDINATED MODELS AND FORMATIVE USER INTERFACE DESIGN

Robert B. Allen

**ABSTRACT**

*Most tools for accessing digitized historical newspapers emphasize relatively simple search; but, as increasing numbers of digitized historical newspapers and other historical resources become available, we can consider much richer modes of interaction with these collections. For instance, users might use exploratory search for looking at larger issues and events such as elections and campaigns or to get a sense of "the texture of the city… how the city was thinking." To take full advantage of rich interface tools, the content of the newspapers needs to be described systematically and accurately. Moreover, collections of multiple newspapers need to be richly cross-indexed across titles and even with historical resources beyond the newspapers.*

**Keywords:** History, Interviews, Modeling Events, Text Processing, User Interfaces

## INTRODUCTION

Because an increasing number of digitized full-text historical newspapers is now available, we can begin to shift from searching them one title at a time to considering the way access to several titles can be coordinated. For instance, in the time frame 1880 to 1920 for Washington DC about ten different titles are available. Similarly, at the state level newspapers from several cities are being digitized and there should be synergies among them.

Collections such as the LC/NEH National Digital Newspaper Program (NDNP) collection are particularly useful for detailed text processing because they provide public domain OCR along with word coordinates and font identification in the META-ALTO format. However, because of factors such as the poor quality of the originals and the necessity of reproduction from microfilm copies, the OCR is of uneven quality. Moreover, there are additional aspects of the newspapers that are not coded in the ALTO files.

Allen et al. (2008) explored a "pipeline" processing model with a series of stages for creating article-level metadata. One step in this processing is segmentation of the page image. This segmentation is based on the identification of large fonts in the text which indicates a headline and the top of an article. The text of each of the segments was then categorized by genre and topic categories based on the standard developed by the International Press and Telecommunications Council (http://www.iptc.org). The metadata assignment was moderately successful for narrow categories which included highly distinctive terms but it was less successful for categories which depended on

nuanced language processing. Allen and Hall (submitted) explored regular news features which were not captured by the original set of genre categories.

In summary, it is easy to process some categories which are associated with distinctive keywords automatically but many other categories are less accurately processed even though most of the content is highly predictable based on factors such as its location in an issue. The more expert knowledge added the higher the accuracy. However, the amount of digitized newspapers is so large that it does not seem feasible for people, even groups of citizens engaged in collaborative correction, to make all the corrections needed. It seems unlikely that complete corrections can be accomplished by the automated process but automated processing can augment the capabilities of the human beings.

### Text Mining and Modeling History

Text mining can be used to identify patterns in the text which should also be useful for improving the text processing. For instance, Allen et al. (2008) reported finding a seasonal pattern for the word "drought". In fact, such regularities are easy to find and here we present rich data for two additional examples. These are drawn from the *Philadelphia Evening Ledger* from mid-September to December 31, 1914. As shown in Figure 1, occurrences of the term "Thanksgiving" peak at Thanksgiving and mentions of the term "Christmas" peak, unsurprisingly, at Christmas. Moreover, other terms related to the holidays such as "turkey" and "Santa" follow similar patterns.

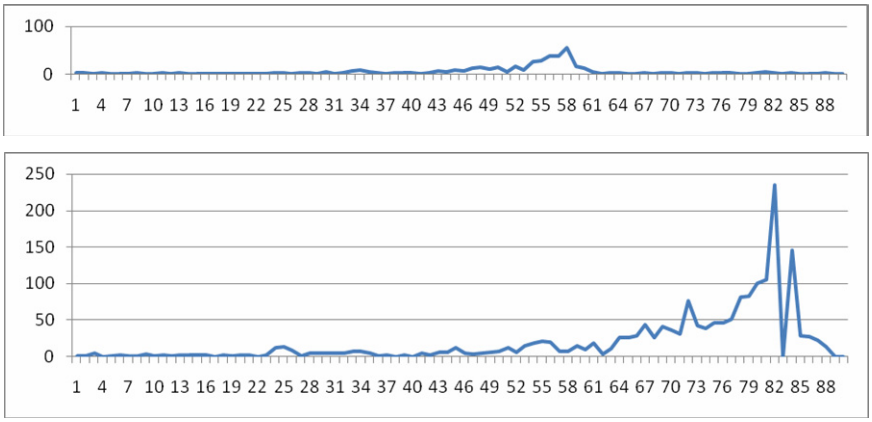


**Figure 1:** Frequencies for the term "Thanksgiving" (upper panel) and "Christmas" (lower panel).

Clearly, there are patterns in the data and they could be useful in making corrections in the text. For instance, they could be used to set weights for OCR corrections. However, to truly understand and exploit them it will be more helpful to develop models which account for and could even predict them. Thus, we want to move from simply observing such regularities to modeling them. Models could be based on many factors. For instance, we would predict very different types of news reports from a rural community than from an urban one. That is, we might develop what we could call "community models" (Allen et al, 2007). These would provide a unified framework for interrelating the people, places, organizations, and events that appear in the newspaper for a town or city.

Local reporting has a context within national and international reporting, even though readers at the time may not have direct access to non-local views. From a historical perspective there is, then, a requirement to enable the correlation of local news to the wider context and indeed to tracking the spread of news in times when local news readers did not have ready access to alternative news sources, such as telephones, television, radio or the Internet. A focus on local context also facilitates access to data that enables understanding of local social networks, social activities, sports, entertainment, community government functioning, advertising norms, and biographic data. These local contexts can then be compared across geography and time to analyze local and regional dispersion of news. In addition, to enhance understanding of the significance of search results, particularly from multiple newspapers, data such as local census data, economic data, or weather data, would be made available for overlaying on the search results. Moreover, relevant named-entities can be derived from many other sources. The "community models" that we propose incorporate these functions. Ultimately, these also need to be combined with newspaper models which encapsulate the editorial, stylistic, and production policies of each newspaper.

In this case, we are particularly interested in determining accessing civic processes, by which we mean government-related activities, which, of course comprise a substantial portion of the news. Figure 2 once again show data from the *Philadelphia Evening Ledger* for the fall of 1914. In this case, the data emphasizes terms related to the election of 1914 which was held on the first Tuesday of November. This is of particular interest because it shows a progression from the campaign to the election. Thus, we have identified a strong sequence of events by examination of word frequencies. This pattern is confirmed with terms such as "candidate" and "rally" for the days leading up to the election and the terms "election", "votes", and "voted" for the days surrounding the election. Obviously, these are very gross measures but they do clearly demonstrate the predictable evolution of events and, thus, we may think of them as being a model of events that generate news.

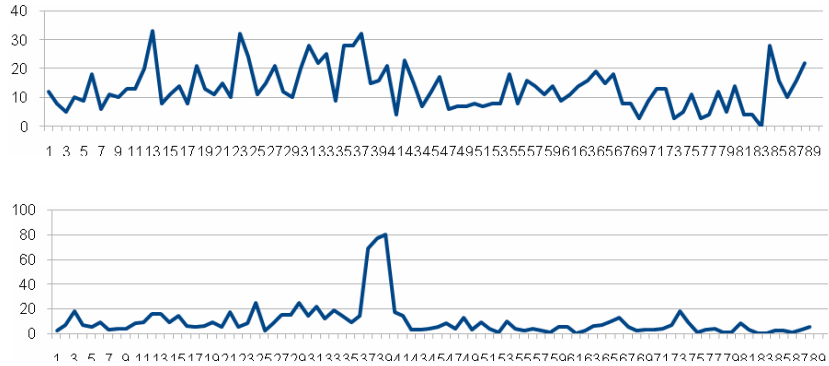


**Figure 2:** Frequency of the term "campaign" (upper panel) and of the word "vote" (lower panel). Note that term "campaign" is highest in the days leading up the election while term "vote" has a peak only for the few days surrounding the election.

**Preliminary Results for Text Mining Concurrent Local Newspapers**

While we have a fairly complete record for significant national events, we have a much less complete record for important events in individual communities. Such a record would be of interest in its own right in addition to being potentially useful for adding constraints for the text processing. However, identifying significant events from individual newspapers can be difficult because of the large amount of material to sift through. One strategy for focusing on significant events would be to compare coverage from different newspapers. There are a few cases where there are multiple newspapers digitized. In particular, the Library of Congress itself has processed a number of historical newspapers for the District of Columbia (DC). Of particular interest, there is a period of about three months in 1906 for which we have digitized samples of two major Washington DC newspapers: the *Washington Times* and the *Washington Herald*.

We collected the OCR output from the first pages of the newspapers during the period of the overlap. We cleaned that OCR by identifying only those words which were also found among the words in the Associated Press and *New York Times* portions of the Linguistic Data Consortium Gigaword corpus. That is, we tried to minimize the OCR errors by comparing the OCR text to words from a very large sample of English text. In addition, a stop list of the 500 most common words in English was also applied. Using the document-frequency measure from information retrieval, we identified terms which showed a striking change in frequency from their overall baseline. Finally, we found those words which showed that change of frequency in both Washington DC newspapers within a three-day window. These words are often indicators of distinctive news stories. Table 1 shows three examples selected from the output of this process.

**Table 1:** Examples of distinctive terms for news stories which appeared across two different newspapers in Washington DC on about the same day in late 1906. This technique allows us to identify news stories of particular significance because they appear in both newspapers.

| | |
|---|---|
| Oct 29 1906 | awful breaking bridge camden coach dempsey drawbridge heroism motorman picked submerged surface survivors thoroughfare trestle windows |
| Nov18 1906 | colon dillon hopes lacking princeton princetons teams tigers yale |
| Dec 31 1906 | ambulances awful belt coaches cotta crowded empty horribly identified mangled relief rescuers splintered takoma terra |

These are preliminary results and much more work needs to be done to make them robust and useful. For instance, the mention of "colon" in the second line refers to the city of Colon, Panama and is unrelated to a main story detected about the defeat of Yale by Princeton in an American football game.

**Toward a Historian's Workbench**

While dedicated researchers can and do exhaustively examine the rich resources of newspapers with microfilm, modern user interfaces for digitized materials should make the job easier for everyone and the barriers to entry much lower for beginning researchers. Thus, it is time to consider how historians might interact

with a much richer set of materials than they have previously been able to do. That is, we might think of developing a historian's workbench (cf., Toms & Flora, 2006). Robert Sieczkiewicz, the Drexel University Archivist, and I are conducting interviews with historians to determine the features historians would find particularly useful (Allen et al. 2010).

One set of issues concerns searching itself. For instance, one historian said. "The *[existing commercial online]* database is good for searching names, but broader topics are hard to research". Another researcher said she used newspapers to fill in gaps in research and corroborate information from other sources. Her exploratory searching included looking at larger issues and events such as elections and campaigns. That is, rather than searching on specific items, she used newspapers to find public opinion about issues such as changes in liquor license laws – to get a sense of "the texture of the city… how the city was thinking". That is difficult to do with simple keyword indexes of the materials.

Another set of issues deals with managing results from searches. One of the historians interviewed said "a log of all searches – this is a huge issue for me". When editing a book manuscript recently, she found it "hugely taxing" to find items she hadn't cited. Similarly, "searches lead to other searches", so she would like ways to see how searches are nested within each other and to return to earlier search results. She also asked for "a visual map telling you where you are in your search" as well a system that lets her easily use multiple windows.

Similar comments are also found in blogs on the Web. "Rachel", who describes herself as a doctoral student in history, presents a set of techniques for searching newspapers.

1. Have a List,
2. Find a thread of some kind,
3. Don't just use one newspaper,
4. Don't fall into the trap of only reading articles that your keywords throw up.
5. Use existing secondary literature,
6. Keep really, really scrupulous notes, and
7. Don't neglect the letters and the advertisements.

Taken together with the interviews, these may suggest that a user interface should have both flexible searching and also tools for annotation and management of the search results.

## CONCLUSIONS

We have presented the value of extending text processing and text mining with modeling of underlying activities that are reported in a newspaper. Indeed, we can view a newspaper as a type of projection of, or perhaps a filter for, the community they are reporting about. In some cases these models can be based on predictable patterns. But, in other cases they are so difficult to predict that it is best to try to identity events as they arise. Thus, we also report results from

comparing the contents of two newspapers from the same city for the same time frame to find stories that are reported in both of them.

While we have explored many attributes of the historical newspapers, they have so much more rich material that we have barely begun to get a complete picture. Moreover, the application of findings such as these will be a huge undertaking. But, the result will be ready access to these exceptionally rich historical resources. Indeed, we envision digitized newspapers cross referenced with a wide range of other historical resources including primary sources as letters and secondary sources such as textbooks. Ideally, they would be woven into an automatically generated historical narrative.

## BIBLIOGRAPHY AND REFERENCES

1. Allen, R.B., & Hall, C. (submitted) Automated Processing of Digitized Historical Newspapers beyond the Article Level: Finding Sections and Regular Features.
2. Allen, R.B., Japzon, A., Achananuparp, P., & Lee, K. (2007). A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. *Human Computer Interaction International,* Beijing.
3. Allen, R.B., Waldstein, I., & Zhu, W. (2008). Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres. In *International Conference on Asian Digital Libraries*, Hanoi, Vietnam, 380-387.
4. Allen, R.B., Zhu, W., & Sieczkiewicz, R. (2010). What to Do With a Million Pages of Digitized Historical Newspapers? Presented at the IConference, Urbana-Champaign IL.
5. Toms, E., & Flora, N. (2006). From Physical to Digital Humanities Library: Designing the Humanities Scholar's Workbench. *Mind Technologies, Humanities Computing, and the Canadian Academic Community*. Edited by R. Siemens and D. Moorman. Calgary: U. Calgary Press, 91-115.