# Automated Processing of Digitized Historical Newspapers beyond the Article Level: Sections and Regular Features

Robert B. Allen and Catherine Hall

The iSchool at Drexel University
3141 Chestnut Street, Philadelphia, PA 19104, USA
{rba,ceh48}@drexel.edu

**Abstract.** Millions of pages of historical newspapers have been digitized but in most cases access to these are supported by only basic search services. We are exploring interactive services for these collections which would be useful for supporting access, including automatic categorization of articles. Such categorization is difficult because of the uneven quality of the OCR text, but there are many clues which can be useful for improving the accuracy of the categorization. Here, we describe observations of several historical newspapers to determine the characteristics of sections. We then explore how to automatically identify those sections and how to detect serialized feature articles which are repeated across days and weeks. The goal is not the introduction of new algorithms but the development of practical and robust techniques. For both analyses we find substantial success for some categories and articles, but others prove very difficult.

**Keywords:** Access, Classification, Digital Humanities, Historian's Workbench, Newspapers, Text Processing.

## 1 Using Structure to Enhance Indexing of Historical Newspapers

Local newspapers are important historical resources. In the past several years vast amounts of historical newspapers, largely created from digitized microfilm, have been generated and we would like to support access to these potentially valuable resources. In the U.S., the major initiative is the National Digital Newspaper Program (NDNP) which is being sponsored by the National Endowment for the Humanities (NEH) and the Library of Congress [1]. Participating states deliver digitized page images and OCR text using the METS/ALTO schema[1]. To date, this OCR has mostly been used for search interfaces[2] but much richer access could be supported by identifying the structure (e.g., articles and sections) of the newspapers. Because such large amounts of historical materials are being digitized, automated methods of processing are needed.

[1] METS/ALTO is a marriage of METS (Metadata Encoding and Transmission Standard) and ALTO (Analyzed Layout and Text Object). The former standard uses XML to encode descriptive, administrative, and structural metadata about digital objects; the latter describes the content and layout of each piece of the digital object.

[2] e.g., http://www.loc.gov/chroniclingamerica/

G. Chowdhury, C. Khoo, and J. Hunter (Eds.): ICADL 2010, LNCS 6102, pp. 91–101, 2010.

The structure of modern newspapers is described by standards such as the International Press and Telecommunications Council (IPTC) family of News Exchange Format Standards (e.g., NEWSML 1, NEWSML-G2, NITF). The IPTC has also developed taxonomies, known as descriptive NewsCodes, for the categorization of newspaper content.[3] There are five sets of these descriptors and we will focus on two of those here. The first of these is Genres, which describes the nature, journalistic or intellectual characteristic of a news object, but not specifically its content. Some examples of Genre include: Background, Daybook, Scener, and Feature. The second taxonomy is Subject Codes which is a hierarchical system to describe content in three different levels of specificity (Topics, SubjectMatter, SubjectDetail). Examples here include: Arts, Culture & Entertainment, Archaeology, and Fire.

Allen et al. [2] reports an initial study of automated processing methods for the OCR from historical newspapers. They first explored automatic methods to segment the pages based on the OCR. Several approaches were tested, such as including semantic coherence among the terms, but for this dataset the best results were obtained from detecting font changes from article headings. This approach was fairly successful overall and especially for relatively large, well-structured articles; however, it was less successful for tightly packed advertisements and notices.

After the articles were segmented, they were assigned to genres with the intention that at least some of the genres (especially news stories) would be processed further (in later research), for instance by assigning subject codes. This sequence, from segmentation to genre assignment to subject code assignment, was described as a pipeline processing model. For the genre and topic categorization, they [2] primarily employed templates based on the presence of specific words. For instance, the weather reports were identified because they included terms such as "temperature", "degree", and "snow". While the OCR text had many errors, there were often enough correct terms for the articles to be identified, especially for those types of articles (such as a weather reports and reports about chess matches) where there were predictable and distinctive terms. In addition to the word templates, matches (exact or partial) for some distinctive phrases (e.g., "Weather Report") were also used as evidence of the appropriate genre category. With considerable tuning, fairly accurate matching was obtained for focused categories. However, the cost of improved accuracy could be a loss of flexibility both when formats and styles change, and across different newspapers. Moreover, it was found to be difficult to maintain the spirit of the pipeline; the second step of determining genre would actually be of use in the first step, determining segmentation [2]. In addition, the IPTC distinction between Genres and Subject Codes did not always match the logic of the pipeline. For instance, advertising is considered by IPTC as type of content. Nor are all genres easy to identify; reviews (literature, theater, music etc) often look no different in format from other news articles.

Rather than pursuing the pipeline processing model per se, it seems more reasonable to frame the task as identifying and using as many of the regularities and constraints as are available to optimize the identification of the various components of

[3] http://www.iptc.org. Other modern newspapers often have XML tags even if they are not IPTC compliant. For instance, in the *Los Angeles Times* has a types of subject category based on "Desks"; for instance, there is a "Book Review Desk". The *New York Times* has "Times Topics".

these newspapers. One constraint which should be useful is that articles on related topics are often positioned near each other. Knowing the section could be helpful for disambiguating OCR and correcting errors because we may then utilize domain specific ontologies which are more successful than general purpose ontologies. As a second type of constraint, we can explore whether we can find regular feature articles. Neither of these tasks is difficult for human analysis but because we need to process millions of pages of historical newspapers there is an enormous advantage to developing automated techniques.

In Section 2, we survey the frequency and stability of sections across newspapers. We then turn to procedures for automatically extracting sections. In Section 3, we extend the subject categorization of based on word counts. We also explore the identification and extraction of regular feature articles.

## 2 Description of Regular Feature Articles and Sections

Surprisingly, there has been little systematic description of the structure and organization of newspapers so we started with that. We found that some feature articles are repeated across days; some fall directly into the genre classifications as defined by IPTC, but others do not readily match those categories. In any event, knowing that those items are repeated and clustered may make them easy to detect. First, we examined what types of clustered and categorized material was typically present in the historical newspapers. Next, we examined how the type of material and its clustering changed across time and how it varies across different newspapers.

We obtained page images from the NDNP team at the Library of Congress for several Washington DC newspapers for the years 1900 to 1910. The *Washington Times* was selected as our primary focus because we had the most complete run for it. We then compared the *Washington Times* for two months during 1904 and then again during 1908. We also compared the *Washington Times* to another Washington newspaper, the *Washington Herald* and a rural Pennsylvania newspaper, the *New Holland Clarion.*

### 2.1 Washington Times, March 1904

We examined each page for the first week of March. Although we kept in mind the general notion of genres, and more specifically IPTC's genre categories, we found it helpful to think in terms of 'sections' as we attempted to identify and explain the pieces of information that come together to make up a newspaper. The backbone of any newspaper is its traditional news articles, those pieces that report on current or recent events, but we also identified a number of other regular items and features:

**Classifieds:** *Times Want Ads* is a dedicated page each day for classified advertisements. Typical categories include; *Help Wanted*, *Situations Wanted, Wanted, For Sale, For Rent, For Hire, Lost, Found, Personal, Miscellaneous.*

**Daybook:** This includes *What Is Going on in Washington*; a list of theaters and show times, excursions etc. Also, a section of possibly paid for advertisements under the heading of *Amusements*, which includes theater and music performances, forthcoming lectures etc.

**Editorial:** Appearing daily with the masthead, a portion of the page is reserved for editorial comment.

**Financial Information:** A daily section which incorporates financial news items, market tables (*Washington Stock Exchange, New York Stock Exchange, New York Cotton Market/ Chicago Grain Market*). Also a section entitled *Current News and Gossip of Interest to Investors*.

**Local News:** In March 1904, *The Washington Times* had regularly appearing features dedicated to the news of regions local to the DC area (Alexandria, Boyds, Georgetown, Hyattsville, and Rockville). Each of these sections may be as short as two or three paragraphs and while *News of Georgetown* appears every day during this particular week, *News from Rockville* only appears once.

**Masthead:** Editorial/publication information including editor's name, office address, subscription prices.

**Notices:** This is a broad category that covers many of the small recurring sections identified from the March 1904 files. Some examples of Notices include Advertised Letters, Church Notices, Death Record, Died, Foreign Mails, Legal Notices, Local Mention, Marriage Licenses, Railroads, Real Estate Transfers, Special Notices, Trustee Sales.

**Poems:** Each day on the same page as the masthead and editorial, the newspaper publishes a short poem or verse.

**Reviews:** Weekly reviews of new theater productions (Tuesday) and literature (Saturday).

**Society News:** A daily section appearing on the same page as the masthead/editorial titled *In the Circle of Society*. It contains news about and of interest to the upper echelons of Washington society - dinners, dances, receptions, people who are in and out of town.

**Sports:** A dedicated sports page appears in the newspaper each day. It incorporates news items, results tables, schedules, etc.

**Weather Report:** Appears on the front page daily. Includes temperature tables and sun rise/set times and tide table.

**Women's Interests:** A significant daily section is *The Home Its Problems and Interests*. Aimed at women, it contains a mix of short features, tips and recommendations on subjects such as fashion, food, children, health and beauty, and haberdashery.

On Sundays, most of the aforementioned features are present, but the paper is substantially longer (about 50 pages) and split into five parts. The first part closely resembles the Monday-Saturday version of the newspaper with a few changes – e.g. more full-page advertisements and the movement of certain features such as society news, financial, classifieds etc to the *Metropolitan Section*. There is also a self-explanatory *Comic Section*, and *Magazine Features,* which includes columnists, special reports and features, and women's fashion. The final part is titled *Colored Section* and contains a mixture of fictional stories and factual features (sometimes pieces of historical interest or sensational true stories), all of which are richly illustrated through photographs and drawings.

### 2.2 Washington Times, November 1904

By November 1904 several sections which we had observed in March had changed names. For example, *The Circle of Society* became *In Society's Circle*, and *Literature* became *In the Book World.* The content of these sections, however, remained largely the same. A more noticeable development is the thinning out of local news sections; *Hyattsville Notes* and *News from Boyds* which appear in the March 1904 papers do not appear in the November 1904 papers we analyzed. Additions to the paper are minor and include a daily cartoon and a section titled *Points in Paragraph*. Both of these appear on masthead/editorial page and the latter appears to be an extension of the editorial, offering pithy comments about current news.

### 2.3 Washington Times, March 1908

By 1908, *Marriage Licenses* and *Died* are combined (along with *Births*) in *Vital Records*. New sections added by this period include the politically focused *What Congress Did* (a short report on bills passed, resolutions adopted, people who spoke in both the Senate and the House) and *Today's Caller's at the White House* (politicians and noteworthy visitors expected that day). Another minor addition is *Court Record*, a notification of the cases being heard in court that day. Changes were also made to the Masthead/Editorial page; *Points in Paragraph* and the cartoon which appeared in November 1904 have not lasted. *News from Rockville* has disappeared, leaving Alexandria and Georgetown as the only local areas to have distinct news sections. Society news enjoys more prominence at this point, moving from the editorial page to earlier in the newspaper. Variously titled (e.g., *Tea and Luncheon Parties*; *Weddings Dinners Teas*) it occupies the majority of a page, sometimes continuing on to a second. Similarly, the women's interests section (now called *Facts and Fads in the Realms of Home and Fashion*), which now also includes *Notes from Stage Folks,* has expanded to fill most of a page.

### 2.4 Washington Herald, 1908

After comparing the *Washington Times* across different years, we then moved to compare it with other newspapers. We first looked at the *Washington Herald*, another daily DC newspaper, founded in October 1906. Looking at the same weeks in March and November 1908, it was easy to identify similarities in the sections and structure of the two papers. Features like weather reports, advertisements, editorial comment, sports and financial sections are expected to exist in both papers, and it is also no surprise that both also contain similar notices – marriage licenses, death record, times of church services, court records etc. Both papers also have sections for women's interests, news from local areas (although the geographic areas covered do differ), and society gossip. Sometimes the similarities extend beyond the content of the paper to the structure itself; the women's section of both papers appears regularly, but not always, on page 7; and the 'Want' advertisements appear usually, but not always, on page 10 of each. While the overall structure is similar, we noted several differences between the papers. Examples of unique features in the *Herald* include readers' letters, fiction serialization and the daily columnist Frederic J. Haskin, who writes on a variety of serious and frivolous topics. The similarities between the papers far outweigh the differences, however, and it is highly likely that an automated process developed to identify sections in one paper could also identify a large number in the other. In February 1939, the two papers merged to become the *Washington Times-Herald.*

### 2.5 New Holland Clarion

As a comparison with the urban Washington DC newspapers, we also evaluated the March and November 1904 and 1908 editions of *The New Holland Clarion*, a weekly newspaper from New Holland, a small rural town in Lancaster County, Pennsylvania. Strikingly, many similarities could be drawn with the *Washington Times*. There were sections for local news (Hatville, Blue Ball, Intercourse, Churchtown, etc.), sports, and finance (although focused on the produce and livestock markets). Generally, these sections were shorter than those in urban newspapers. National news was sparsely reported compared and rarely made the front page.

This is a small community newspaper; a front page story from March 1904 was titled *Many Bones Are Broken* and concerns a number of people who had broken limbs during inclement weather. Like the *Washington Times,* the C*larion* includes a substantial number of advertisements and classifieds, and also a number of notice-type sections including unclaimed letters, times of church services, and railroad timetables. Marriages and deaths are also reported, though sometimes more sensationally than in the *Times*; example headlines include *The Work of the Reaper* and *Gone to the Great Beyond*. *The New Holland Clarion* also has a section devoted to people who in and out of town (*Points Purely Personal)* and, in November 1908, it introduced a women's interest feature called *Home Circle Department*, aimed at "tired mothers as they join the home circle and eveningtide".

### 2.6 Summary of Observations

The amount of consistency across the three papers is notable. Structure exists that should be easily recognizable by both human and automatic extraction methods. We know that certain features such as sports, classifieds, financial, society news and women's interests appear in the *Washington Times* daily. We also know that when an edition of the paper is 12 pages long, the sports section is likely to appear on page 8, the financial information on page 9, and the *Times Wants Ads* on page 10. In addition, certain features of the *Washington Times* are likely to appear together; the masthead and editorial comment are always on the same page, and the society news regularly appears with them.

However, there is also considerable change in sections over time and across newspapers. Indeed, not only do the details of the sections change but the conceptual organization itself changes. For instance, Vital Records may or may not appear as a separate section and when it does it appear typically includes a mix of Births, Deaths, Marriages.

## 3 Automated Procedures

So many pages of historical newspaper have been digitized that is not realistic for them to be manually marked up. Thus, we explore techniques for automatically identifying the sections and features described in Section 2.

### 3.1 Test Data Set

Because sections change frequently, we cannot simply create a template to find them consistently; instead, we need to develop automated procedures for finding the sections. We focused on five categories of sections which we judged to be fairly robust and which, if correctly identified, would account for a substantial portion of the newspapers. For each of those sections, we manually listed the pages on which they appeared for each day during March 1904 and established rules for the coding. On some days some sections filled more than one page, for instance, sports and classified advertising sometimes covered one-and-a-half pages. In those cases, we coded only one full page. However, when these sections covered less than one page they were coded for that page and, indeed, these could be two sections on one page.

### 3.2 Section Identification from Tagged Articles

The first technique we explored was based on the article-categorization approach of Allen et al.[2]. If we found a sufficient number of sports articles on a page, we would conclude that that page was a Sports section. The accuracy for this procedure was low, even for distinct sections such like Sports. Apparently, there was enough error in the article-level identifications that accurate categorizations could not be made with this approach and we did not pursue it.

### 3.3 Section Identification from Page-Level Word Lists

Rather than focusing on articles, we shifted to considering entire pages using a word-counting technique similar to the method in [2]. Specifically, we developed word lists for each of the five types of sections on which we were focusing. We then found the average frequency for each of the terms across all the pages for that month. Next, we compared the frequencies separately for each page to the frequency for the entire month. If the page frequency exceeded the overall monthly frequency by a large multiplier (e.g., 30 times), that was considered to be a hit. Then, if a minimum number of such matches (e.g., 4) were obtained for a given category we identified it as that type of section.

We applied this method to the 324 weekday pages of the *Washington Times* for March 1904 for which complete data were available and we compared the results with the section coding from Section 3.1. Table 1 summarizes the results with the page-level word-list technique. This technique turned out to be quite successful for the sections on which we focused but less successful when we explored other types of sections such as Editorials. Although Sports had many distinctive terms, we found only a few distinctive words which were associated with Editorials and even those were not reliably captured by the OCR. The lowest accuracy observed in Table 1 was for Society. While that section often included news about galas with royalty and diplomats, there were some days when the events reported were more mundane and hence more difficult to distinguish from other news.

**Table 1.** Hit and False Alarm Ratios for several sections for the weekdays (324 pages) during March 1904

| Type of Section | Hit Ratio | False Alarm Ratio[4] |
|---|---|---|
| Classified Advertisements | 1.00 | 0.00 |
| Home and Family | 0.96 | 0.65 |
| Society | 0.62 | 0.00 |
| Sports | 1.00 | 0.53 |
| Stocks and Finance | 0.92 | 0.00 |

Our strategy was to select methods that would minimize the need for human intervention in the categorization process. However, it is unlikely that human intervention can be entirely eliminated and the goal should be to find a balance in which human intelligence can be applied with the greatest leverage. In an early run, we noted that the Home and Family category was often falsely identified on a page of classified advertisements because there were many domestic terms among the classifieds. To minimize those false alarms we instituted a rule such that if the page is recognized as having Classified Advertising, then it should not also trigger the Home and Family category. In any event, these do not seem to be very serious errors as the terms for the two categories overlap a lot. Of course, there were some true errors for instance, when a cluster of Finance stories appeared on the front page and that cluster was identified as a section.

Because the five sections we studied were fairly standard and broad we expect the results could likely be generalized across time and to other urban newspapers. However, as noted in Section 2.4, the Finance section for the rural *New Holland Clarion* was quite different from the Finance section for the urban newspapers. Finally, although the analysis in Table 1 excluded Sundays because of their very different structure of sections, when we did look at the Sunday newspapers we found that their Sports and Finance sections were several pages long and these were consistently correctly identified.

### 3.4 Identifying Regular Features

Some of the items described in Section 2 above were included because they were clearly demarcated in the newspaper as regular features with distinctive and repeating headings (e.g., *News of Georgetown*). While many of the larger sections had such headings (e.g., *News and Gossip of the Day from the World of Sports*) the shorter items might be better called regular features rather than sections. They remained distinctive in comparison to traditional news stories which almost always had unique titles. In any event, identifying the items with repeated titles should help us to parse the sections of the newspapers. Moreover, because the strings of characters were so distinctive and were repeated so often, it seemed promising to identify them with our basic text processing tools. Furthermore, because the strings were generally several words long identifying them would also be robust to OCR errors.

[4] A maximum of one false alarm was counted per day.

For each article during the month of March 1904, we extracted the first line of articles which contained more than five characters. We then compared those to the first lines of text from all other articles extracted for that month. Of course, the matching was complicated by OCR errors so partial matches were based on counting the number of overlapping letters in two strings and the order of the characters was ignored. Article matches were counted if at least 85% characters matched. Out of 9556 items, 667 sets of matching articles were identified. The majority of the matches were for advertisements being run for several days. It appeared that short articles with repeated headings were advertisements. This suggests a simple technique for detecting certain types of advertisements and that is also useful for identifying the contents of the historical newspapers.

All of the feature titles described above in Section 2.1, with the exception of "*What is Going on in Washington*" were found using this method. However, in some cases because the titles were split across lines, only the first words were matched. There were also some limitations for the automatic methods. For instance, many datelines (e.g., "NEW YORK") were observed though, presumably, those could be filtered out. Poems were not detected because the lead had only the title of the poem. There was, however, distinct separation lines demarcating poems and conceivably those separators could be used to find the poem text.

The partial match procedure can also be used if the goal is not discovery of items which are frequently repeated, but finding instances of titles which are known to be in the text. For instance, the title of the sports section was "*News and Gossip of the Day from the World of Sports*" but it was not associated with a specific article. In the same vein, using the banner heading for the Society category, "*In the Circle of Society*" could be used to reduce the errors for that category in the word-based categorization technique as describe above.

## 4 Discussion

The techniques described in this paper should help the automated indexing of the historical newspapers and, thus, ultimately support better access for users. Ultimately, we believe that the newspapers could be a substantial component for a Historian's Workbench [3, 4]. More immediately, the complete index would need to be compiled and even with a great deal of automated processing, it seems likely that some level of human intervention will be needed. Thus, we envision an interface and content management system for managing updates. Perhaps this could be a version of the "collaborative correction" technique that was originally developed by the National Library of Australia [5].

We believe that automatic methods should be used to leverage human input and there are several promising areas to explore. The categorization techniques described here are straightforward and robust but are probably not as sensitive as techniques which systematically assess the most predictive terms for each category. Importantly, we might be able to address the problem of sections shifting across months. While there some variability of sections from day-to-day, those shifts are generally transient, and a high-level program could examine output across weeks to learn the prototypical pattern. Such a program might detect and flag shifts in editorial policy for sections (e.g., when a new section is added or an old one is dropped). A further interesting challenge presented by the shifting of sections is how to describe the contents of the newspaper with structural metadata systems such as METS.[5]

Data management itself can be a substantial challenge. While the data sets used here are already very large, inferences such as those based on imperfect knowledge bases and on collaborative correction may multiply the amount of errors. Inevitably some incorrect information is likely to creep into the knowledgebase. Even without errors there will be ambiguous (e.g., which Mr. Smith was mentioned in a given article) and disputed information, so the provenance of all inferences should be maintained indicating what updates were done and why they were done.

The contents of newspapers are clearly structured; yet those structures are sometimes fuzzy and flexible. At the item-level, it is difficult to define what constitutes an article [2]. In this paper we have focused only on sections such as Sports which were relatively unambiguous while acknowledging there are several other content clusters. We should consider ways to accept and even embrace that variability. We could consider articles and sections as prototypes rather than as absolute categories. We might consider the faceting of metadata attributes or even consider degrees of applicability for them. Another way to address these issues is to consider the units of the newspaper to be genres[6] [6]. Indeed, the interrelationship of the elements could be explained by considering them as “ecology of genres” [7]. Thus, we might find that when one section changes that there is a realignment of other sections to match it. Indeed, such shifts might be found across newspapers when several papers serve a community.

While there are limitations in the ability to identify and even the define articles and sections, it should also be emphasized that we can do this well most of the time. Furthermore, there is an increasing synergy among the pieces as we pin down more of them. For example, identifying sports sections should help us to identify sports articles which appear in those sections. It could even help us to disambiguate the OCR text. There are also many factors to consider beyond those that we have considered thus far. For instance, when doing the analyses described above, we observed that the classified advertisements consistently had the highest word count of any other pages. A further synergy may be created by identifying the type of events being described. Some of our other work [4] has begun to explore developing community event models by combining evidence from several different types of historical resources. Furthermore, identifying events may allow us to develop visualization interfaces such as timelines [8] and, ultimately, to improve public awareness and understanding of history.

**Acknowledgments.** Catherine Hall was supported with an IMLS LC21 Fellowship to the College of Information Science and Technology of Drexel University. Other support was provided by a NEH Digital Humanities Start-up Grant to Robert B. Allen. We thank Ray Murray of LC for the digitized newspapers from Washington DC. We also thank the Lancaster County Historical Society and Access Pennsylvania for the *New Holland Clarion*.

[5] While METS/ALTO is focused on describing OCR, we would also like to develop structural descriptions for the sections. However, as we have seen, it is sometimes difficult even to define sections.

[6] This is a broader sense of “genre” than the IPTC Genre codes discussed in Sections 1 and 2.

## References

1. Murray, R.L.: Toward a Metadata Standard for Digitized Historical Newspapers. In: Proceedings of IEEE/ACM JCDL, pp. 330–331 (2005)
2. Allen, R.B., Waldstein, I., Zhu, W.Z.: Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres. In: Buchanan, G., Masoodian, M., Cunningham, S.J. (eds.) ICADL 2008. LNCS, vol. 5362, pp. 380–387. Springer, Heidelberg (2008)
3. Toms, E., Flora, N.: From Physical to Digital Humanities Library: Designing the Humanities Scholar's Workbench. In: Siemens, R., Moorman, D. (eds.) Mind Technologies, Humanities Computing, and the Canadian Academic Community, pp. 91–115. U. Calgary Press, Calgary (2006)
4. Allen, R.B.: Improving Access to Digitized Historical Newspapers with Text Mining, Coordinated Models, and Formative User Interface Design. In: IFLA International Newspaper Conference: Digital Preservation and Access to News and Views, pp. 54–59 (2010)
5. Holley, R.: How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine 15(3/4) (March/April 2009)
6. Ihlström, C., Åkesson, M.: Genre Characteristics – A Front Page Analysis of 85 Swedish Online Newspapers. In: Proceedings of the Proceedings of the Hawaii International Conference on System Sciences (2004)
7. Foulger, D.: Medium as an Ecology of Genre: Integrating Media Theory and Genre Theory. Media Ecology Association (2006)
8. Allen, R.B., Nalluru, S.: Exploring History with Narrative Timelines. In: Smith, M.J., Salvendy, G. (eds.) HCII 2009. LNCS, vol. 5617, pp. 333–338. Springer, Heidelberg (2009)