News Media Section

Coordinating Newspaper Digitisation: Some Facts and Figures

Hans-Jörg Lieder
Staatsbibliothek zu Berlin, Berlin, Germany.
hans-joerg.lieder@sbb.spk-berlin.de

Abstract:

Numerous newspaper digitisation programmes and projects have taken place in recent years. While many of these projects were exclusively undertaken by individual libraries, some digitisation activities were and are part of a greater scheme that requires some sort of coordination among the involved institutions.
This paper will look at the example of Germany, where a national digitisation scheme for historic newspapers is currently in development. The German National Union Catalogue of Serials provides the required factual data basis for designing such a national scheme. This data provides valuable insights in the particularities of newspaper publications over time, in as far as these are kept in libraries. The evaluation of the data will be centred around three major questions: To which extent are historic newspapers available in libraries? What do we know about the regional distribution of historic newspaper publishing? In which way are historical newspapers distributed in libraries?
Preliminary answers to the above questions allow for some practical recommendations to be presented as a conclusion of this paper.

Keywords: Historic newspapers, Digitisation, Statistics

Introduction

Historic newspapers, being a reflection of a wide range of information and public debates, are a material type that is interesting for almost everyone, for scholars and the greater public alike. However, in the pre-digital era it was nearly impossible for those with a historic interest of some kind to consider newspapers on a large scale, mainly for practical reasons. The availability of large digital repositories of full text newspaper content therefore provides, for the first time, a suitable mechanism to create and gather new historical insights – insights that were impossible to gain without the consideration of fast-paced public communication as we find in newspapers, particularly of the 19th and 20th centuries.

When thinking about a national digitisation scheme for newspapers under the auspices of the German Research Foundation (Deutsche Forschungsgemeinschaft – DFG)[1], one specific user group, scholars, was in particular focus. These were asked as to what they would wish for with respect to the digitisation of historic newspapers that are kept in libraries. Not surprisingly, it was said that a national digitisation scheme should in principle represent the entire chronological and regional spectrum of the material type. More specifically, digitising libraries should consider the newspapers of major national and regional urban centres, but also of rural areas, newspapers with a long life span, the ones that were reputable and innovative even if only short-lived, those that had a specific thematic focus and those with a general scope, and finally those that were representative of the entire political spectrum of the country or any regional precursor – in short: everything. Add to this a multitude of use cases from the greater public, from hobby genealogists to readers that are interested in any aspect of regional or local history, and the libraries' task of selecting newspapers for digitisation appears to be insolvable. Striving for completeness could be one “easy answer”. However, even if libraries aim for comprehensive digitisation, they will still have to select titles in order to prioritise their activities, based on selection criteria that will differ from context to context.

The ZDB: Content and caveat

The Union Catalogue of Serials (Zeitschriftendatenbank – ZDB)[2] details the holdings – serial publications only, among them newspapers – of around 4.200 institutions, mainly libraries, in Germany and Austria. While public libraries are, by and large, excluded from this number, it includes all major and university libraries. Archival holdings are represented on a much smaller scale.[3] In total the ZDB contains around 1,8 million bibliographic title records and more than 15 million records that describe local holdings and their availability. Around 60,000 of the title records describe newspapers. The following observations are therefore not an overview of the newspaper publishing history in Germany but rather provide an insight into collection building policies of libraries in Germany with respect to newspapers.

With regard to the definition of “newspaper” and “newspaper title” this paper assumes a formal position, and simply considers librarian classifications: Newspapers are those resources that were catalogued as newspapers by librarians. It has to be noted that librarian definitions may vary within one country, from country to country and, outside of libraries, from one scholarly subject to the other. The definition of a “newspaper title” follows the same logic in the context of this paper. It is simply any entity that is considered to be a newspaper title by librarians in Germany and Austria. Librarians in other countries may follow different rules which may well result in a different definition of a “newspaper title”.[4]

[1] http://dfg.de/.

[2] See http://beta.zdb-opac.de/zdb/index.xhtml and http://www.zeitschriftendatenbank.de/startseite/ (in German only).

[3] In the following the institutions participating in the ZDB will simply be referred to as “libraries”. At the same time it is to be remembered that the ZDB is a material-specific, not an institution-specific undertaking, that engages institutions across the sectors.

[4] Many newspapers have predecessors and successors, i.e. a newspaper may change its name and this changed name may or may not result in librarians classifying the new name as a new „newspaper title“. An example: the „Neue Leipziger gelehrte Zeitungen“ is listed in the ZDB as having the predecessor „Neuer Zeitungen von Gelehrten Sachen auf das Jahr ... Theil“ and the successor „Neue Leipziger gelehrte Anzeigen“, i.e. three titles in total. Under a different set of rules these 3 newspaper titles could constitute only one or two titles, or even more than three.

Obviously, library data is subject to change and, in some cases, error. The ongoing cataloguing activities of 4.200 institutions in the ZDB assure constant change of the data set. However, both the frequency of changes and updates, and the error margin of the ZDB data, that cannot be detailed but must be assumed to exist in spite of great editorial care, are not thought to jeopardise the overall findings of this paper. It needs to be clearly stated though, that all numbers presented in the following should be rather seen as being indicative of trends, and not as exact descriptions of an unchanging situation.

The data sample

In total the ZDB describes just over 60,000[5] newspaper titles from all over the world that are kept in German libraries. For the specific purpose of this paper, a subset of these newspapers had to be selected. The following selection criteria were applied:

- material type = newspaper
- date of first publication (year only) = 1500-1944[6]
- form of publication = print
- place of publication = Germany AND German Empire

OR

- language of publication = German

When these selection criteria are applied, the total of 60,000 newspapers can be narrowed down to some 22,000 titles that were published between 1600 and 1944 and either originated in Germany and its regional precursors, or were published abroad, but written in German. These approximately 22,000 newspapers are the subject of the following analysis.

Language and publication frequencies

In the light of the communicative nature of newspapers it comes as no surprise that around 95% of all newspapers under consideration were written in German. However, all in all we find 40 different languages in the sample, the most frequent foreign language being French, followed by Polish and English. It is interesting to note, that Latin, the major language of publication in Germany until roughly 1750, only ranks as number 14 in newspapers. Seven languages are represented with one newspaper title only; among these are rather exotic ones like Sorbian, Ottoman, and Esperanto. This multitude of languages clearly shows the richness of the German newspaper tradition.

Another interesting aspect concerns the publication frequency of newspapers. Based on the expectations derived from modern newspapers, one might assume that daily newspapers are the rule, and larger publication intervals the exception. However, throughout the centuries a weekly publication interval is the most frequent case, followed by daily newspapers, and, on a much smaller scale, other publication intervals known to the ZDB.[7] The ratio of daily newspapers increases significantly as time advances, but absolute numbers never reach those with a weekly publication interval.

[5] In the following only rounded figures will be provided.

[6] The cut-off year of 1944 was chosen because it marks a clear turning point in German history, but also because the year is near the copyright threshold.

[7] These publication intervals are: 3-5 times per week, twice per week, bi-weekly, half-monthly, monthly, irregularly.

The special case of early newspapers: 16th and 17th centuries

It is a widespread assumption that the first newspaper in modern Europe was published in Germany in 1605.[8] However, the formation of the publication type "newspaper" in Europe is not undisputed. The types of publications of the 16th century that are sometimes referred to as newspapers, usually when talking from a newspaper-centric point of view, are more often than not kept in libraries' departments of rare books and not really thought of as newspapers by their curators. In the current context it is without merit to discuss whether these early examples are to be considered newspapers or not. Because the ZDB does not contain enough data about relevant 16th century publications, a meaningful analysis simply cannot be undertaken.

During the past years, the State and University Library Bremen has digitised all existing German newspapers of the 17th century – and some of the 16th century – with a total of 375,000 pages.[9] This project produced interesting results with respect to the first centuries of newspaper publication in Germany. As of April 2016 with Bremen not yet having completely added their digitisation data to the ZDB, the ZDB lists just over 120 newspaper titles of the 16th and 17th centuries, while Bremen reported to have digitised around 800 titles of the same period. One main reason, which is specific to the early centuries of newspapers and/or newspaper-like publications, accounts for this massive discrepancy: Librarians in Bremen could fall back on the decade long work of the University's Institute for Press Research (Institut Deutsche Presseforschung) that collects microforms of early newspapers. These early newspapers were clearly identified by the Institute's specialists with criteria that followed a press related logic and not a formal librarian rationale. This explains why a lot of the data in question currently is part of catalogues or databases other than the ZDB. Eventually Bremen's newspaper data will be entered into the ZDB and that will complete the currently sketchy data for the 17th century.

Numbers are more reliable for the 18th to 20th centuries and the following figure, detailing the first publication of newspaper titles and their publication frequencies, is therefore restricted to that time period.[10]

[8] See e.g. https://en.wikipedia.org/wiki/Newspaper#Europe or https://de.wikipedia.org/wiki/Zeitung#Geschichte. The newspaper title mentioned there is the „Relation aller Fürnemmen und gedenckwürdigen Historien", printed in 1605 by Johann Carolus in Straßbourg.

[9] See http://brema.suub.uni-bremen.de/zeitungen17 (German only). The claim to have digitised all newspapers of the 17th century obviously needs to be taken *cum grano salis*. Librarians in Bremen are aware that parts of the tradition may have been overlooked and that missing items may surface in other libraries.

[10] Numbers for the period 1945-1999 were added in Fig. 1 to show the general development of the newspaper tradition, but are not considered in the further analysis.

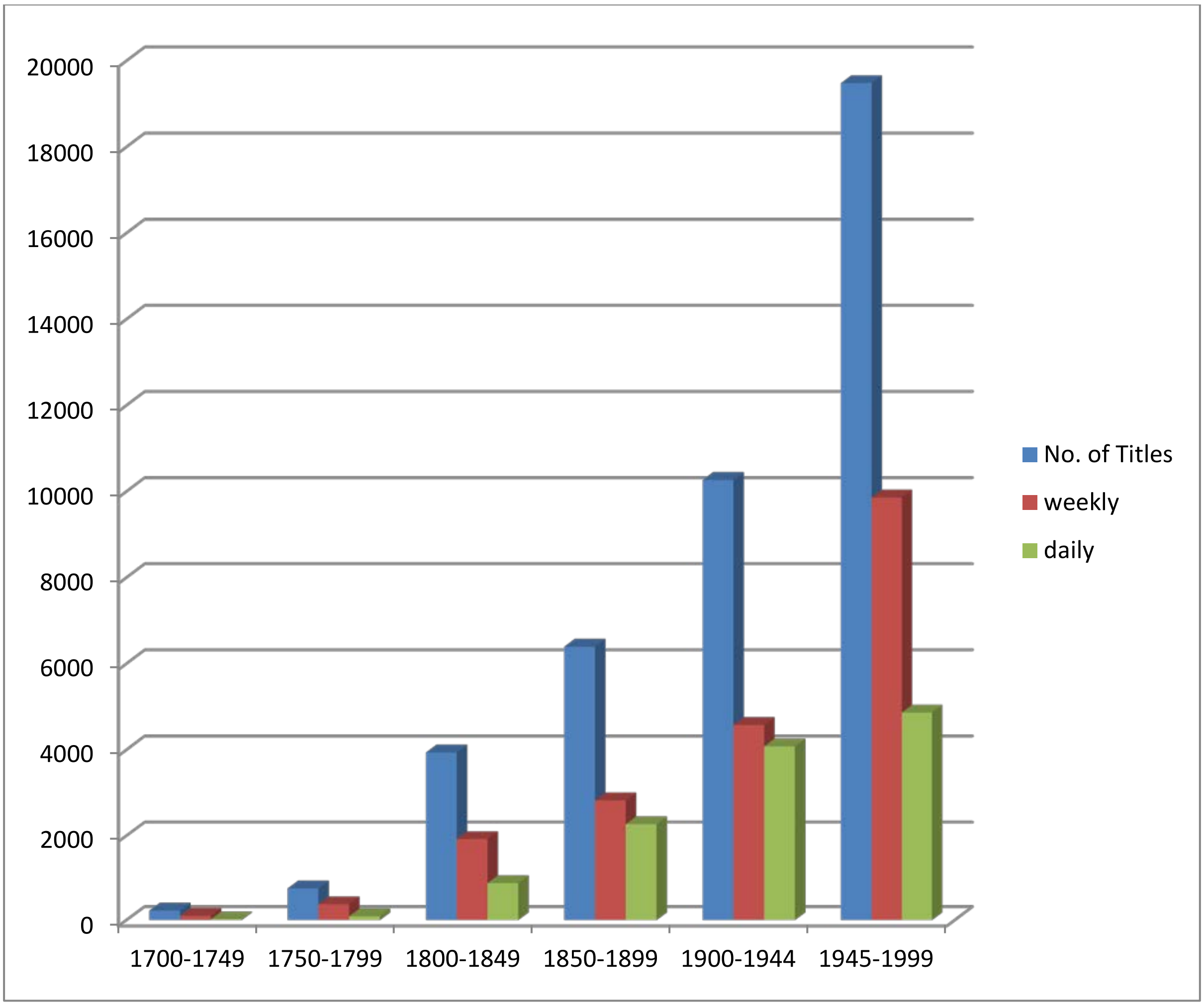

Figure 1: Number of printed newspaper titles being published for the first time per half century, and common publication frequencies.

It may be assumed that the numbers for the 18th century are somewhat debatable due to the same typological reasons, but on a smaller scale as is the case in earlier centuries of newspapers. However, the general tendency over time is clear: numbers of newspapers have grown constantly and, particularly since the 19th century, where figures easily pass the 4-digit threshold, significantly until the end of the 20th century.

While Fig. 1 conveys a picture of the numbers of newly established newspapers, it does not say anything about the longevity of the individual newspaper title. The life span of any newspaper may actually be constricted to one issue only or, to the other extreme, may cover entire centuries. For a better understanding of the actual life-spans of newspapers a number of values are considered in the following table.

time period	mean	median	standard deviation	minimum	maximum
1700-1749	28.28	9	42.82	1	237
1750-1799	18.36	5	31.09	1	213
1800-1849	17.60	4	28.90	1	205
1850-1899	22.39	9	26.48	1	156
1900-1944	8.23	4	10.89	1	101

Table 1: Life-span of newspapers (all values in years)

The data set is clearly widely spread out, but less so as time advances, as can be seen best by the standard deviation values. Throughout the centuries a large part of the newspapers had relatively short life-spans as is indicated by the median values.[11] The maximum values, i.e. the maximum number of years a newspaper existed, are all the more notable, particularly since even as early as in the 18th century newspapers were created that existed for well over two centuries. The time period of 1900-1944 shows that the press market in the first half of the 20th century was mainly shaped by the turmoils of two World Wars and political changes that stretched from the Empire to fascism in less than 50 years. In this case the largest number of new newspapers per half-century corresponds to extremely short life-spans of most individual titles.

Places of publication

Of course, the ZDB data also provides information about places of publication. In the present context a more detailed look at that data facet is worthwhile, because some assumptions regarding the availability of those newspapers in libraries today may be deduced. The capitol Berlin is by far the most frequent place of publication, though not clearly leading in early years, but increasingly so after the foundation of the German Empire in 1871, with a total of nearly 1,900 titles. Other urban centres follow, the 10 most important ones are, in descending order: Munich, Vienna, Cologne, Leipzig, Hamburg, Nürnberg, Dresden, Stuttgart, and Augsburg, where still more than 200 newspapers were published. Together these 10 cities roughly account for a quarter of all newspaper titles. It may be assumed that a relatively large part of these urban newspapers were of more than just local significance and were therefore fairly widely distributed at the time. Furthermore, the urban centres were, throughout modern history, home of well established archiving institutions, available for the long-time storage of printed materials. It seems safe to conclude, that these two factors combined have ensured that a considerable part of these urban newspapers are available in many libraries and easily available for digitisation today.

The above is applicable to a quarter of all newspapers – the situation for the remaining 75% is by far more complex: overall more than 3,500 places of publication are named, of which roughly 1,700 are only listed with 1 title.[12] These numbers clearly indicate the regional diversity of newspaper publishing in Germany. Such regional diversity also suggests that many of these newspapers would have been of local significance only and are available today

[11] This appears to be also true for newspapers of earlier times, as is indicated by some snap samples taken from the Bremen data.

[12] Due to orthographic variants in the spelling of place names (e.g. Frankfurt a.M., Frankfurt/M., Frankfurt/Main, Francfort) actual numbers will be significantly lower. It is not to be expected though, that this would affect the ratio of numbers significantly.

in few or even only one library. An overall look at the regional distribution of newspaper publishing in Germany without consideration of the Top 10 places of publication named above results in the following figures:

no. of titles per place	**no. of places**	**no. of titles per place**	**no. of places**
101-200	14	21-30	38
51-100	28	11-20	166
41-50	17	6-10	328
31-40	22	1-5	2,998

Table 2: numbers of publication places and published titles per place of publication.

Holding institutions

So far information about the actual newspapers has been presented – numbers, life-spans, frequencies, languages, geographical origin – in order to gain a better understanding of the task at hand, and an approximation of the sheer size and diversity of the material. In the context of designing coordinated digitisation schemes, it is equally important to review data that concerns those institutions that keep newspaper holdings.

More than 1,650 institutions have one or more of the 22,000 newspaper titles under consideration in their holdings. It has to be noted that usually only parts of the published issues of any newspaper title are available in any specific library. Complete holdings of entire life-spans at least of long-running newspapers are the exception, not the rule.

Again, it is not surprising that the largest library of the country, the Berlin State Library, owns by far the largest newspaper collection with more than 6,500 titles. In fact, 8 of the 10 institutions with the largest newspaper collections measured by number of titles are seated in one of the top 10 places of newspaper publication mentioned above, which indicates a solid tradition of newspaper archiving in these urban centres.

time period	**no. of libraries with holdings**	**maximum no. of titles per library (median in brackets)**	**no. of libraries with 1 title only**
1700-1749	240	97 (3)	17
1750-1799	509	275 (3)	162
1800-1849	781	980 (3)	196
1850-1899	1,258	1,383 (3)	394
1900-1944	1,248	3,006 (3)	422

Table 3: Number of libraries with relevant holdings per time period, maximum number of newspaper titles per library, number of libraries with one title only.

All numbers are simply related: the larger the number of published newspapers, the larger the number of institutions with holdings, and the other numbers increase accordingly over time. Throughout the entire time period only 3 libraries are leading in number of titles.

The specific newspaper titles that are available in the largest number of institutions – the most popular newspaper title, the “Bauwelt”[13], is available in more than 250 libraries – mostly date back to the second half of the 19th and the first half of the 20th century. Among the 50 most widely available titles – libraries per title range from almost 300 to 80 – 10 titles data back to the 18th century. A look at the other side of the spectrum, at those newspaper titles that are available in one library only, reveals astonishing numbers: Around one third of all newspapers, well over 7,000 titles belong to this category. If these newspapers were to be digitised, only one library could actually provide the resources.

Work already done

The ZDB also specifies whether a title, or part of a title, has already been digitised. About 2,300 of the total of 22,000 newspaper titles have been worked on in the past. Though it may seem as if well over 10% of the work has already been done, this figure is in need of correction. It is evident, that mostly newspapers with a rather short life-span have been completely digitised. For titles that span larger time periods, regrettably often only small parts are available in digital form. Two typical examples taken from the actual data:

Title 1, published: 1808-1825; 1839-1936; digitised: 1871-1880; 1894-1933;
Title 2, published: 1852-1918; digitised: 1864-1865; 1868-1969;

Whatever the motivation behind such seemingly random selection may have been, from a panoramic point of view such activities appear almost futile and result in nothing more than the presentation of random fragments in local digital libraries. The actual percentage of digitised content therefore has to be lowered and would range significantly below the 10% threshold.

A stab at the impossible: a word about quantities

With regard to quantities as expressed in pages, the above analysis has made clear that without closer and more extensive scrutiny, much of which has to consider the originals rather than catalogue data, it is nearly impossible to specify or even vaguely speculate on actual numbers of pages. Calculations that are based on the numbers of titles and pages in the Austrian ANNO newspaper portal[14] indicate a total of almost 300 million pages of German historic newspapers for the period 1600 to 1944. Written recommendations to the Federal Government Commissioner for Culture and the Media, and to the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany that were published in 2015[15] and based on actual surveys, reported another 75,000 metres of newspaper holdings to exist in archives. If we suppose that this overall figure relates to the time period under consideration in a comparable way to the ZDB data, another 25,000 metres of historic newspapers could be added to the library calculation.
However, it cannot be stressed enough that all of these figures should be taken as estimates rather than precise calculations. The exact extent of available newspaper holdings in libraries simply cannot be specified with catalogue data only.

[13] See http://beta.zdb-opac.de/zdb/title.xhtml?idn=011453192&mf=nonlatin.
[14] http://anno.onb.ac.at/.
[15] The recommendations were provided by the Coordination Office for the Preservation of the Written Cultural Heritage at the Prussian Cultural Heritage Foundation.

Conclusion

The sheer quantities of German historic newspapers kept in the libraries of the country make two things obvious right from the beginning: only many will achieve the goal, and this will take a lot of time. It will therefore be necessary to prioritise the activities based on widely accepted criteria. Many user groups and stakeholders and their specific use cases should be considered – academia, the greater public, but also businesses that may create content or services from and around historic newspaper content. Selection criteria will by necessity also consider the actual content, legal and conservational aspects, as well as organisational factors – usually a mix of all.

The data gives the clear indication that cooperation is without alternative. Much can be achieved by relatively few large libraries. However, a well-conceived digitisation scheme needs to involve a much larger number of libraries, particularly for the completion of title specific holdings and for the digitisation of rare newspapers. It is the author's firm belief that the need for cooperation should be seen as an opportunity rather than a burden. For many smaller institutions such cooperation could result in the building of much required digitisation expertise, enabling them to integrate digital workflows in their daily work routine.[16]

In the light of the complexities of the material it is advisable to maintain a monitored roadmap or registry of some kind. It was noted that there are uncertainties as to the very definition of what a newspaper is, and that therefore catalogue data and digitisation results may be “hidden” in other general or specific library catalogues. The exact definition of a newspaper title may also be subject to debate, and the periodicity of newspapers adds another layer of complexity. Supplements to newspapers, so far not even mentioned, also require particular attention, since these are sometimes considered as almost independent works. There are examples of libraries having digitised a supplementary publication but not the newspaper it came along with. Identical supplements may be added to a number of newspapers, which also makes close monitoring inevitable.

An instrument like the ZDB is ideally suited to maintain and monitor the data basis and progress made. If such an instrument is not available, the creation of functional equivalents should be considered.

The large quantities at hand also suggest, that progress will not be made in due time if time-saving digitisation from microfilms is not considered to be a reasonable option in many, if not most cases. Experience has shown that digitisation from quality microfilms leads to OCR[17] results that are comparable to the results of digitisation from paper originals.[18]

The wide distribution of small parts and fragments of existing digitised historical newspapers in a number of digital libraries is a nuisance and a threat to the use potential of the material. For many use cases the size of the newspaper collection that is readily available for searches is the most important factor. The aggregation of digital full text content is therefore one of the much required features wanted by users. Such aggregation can take place at a regional or

[16] This is not a plea for in-house digitisation. Understanding digital workflows may very well result in the decision to out-source digitisation.

[17] Optical Character Recognition.

[18] The Europeana Newspapers project has demonstrated that OCR based on images that were scanned from microfilms resulted in Bag of Words success rates of over 84% for German. For details see: http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D3.5_Performance_Evaluation_Report_1.0.pdf.

national level[19], and even on a European or global scale. Obviously, the allocation of license statements to the digital resources that allow for as many uses as possible is a necessary precondition for that. Users also frequently require big chunks of data for their work, e.g. for applying text mining techniques in the context of Digital Humanities. Again, this wish calls for open licenses, but also for suitable interfaces and data access options to be available to the users of the digital collections of historical newspapers.

The data inspected here is that of libraries owning German historic newspaper collections. Though it is clear that newspaper traditions differ widely across Europe, it is to be hoped that some of the findings and conclusions are, to some extent, representative for other geographical contexts.

Finally, this paper has inspected and analysed exclusively bibliographic data. In doing so, only some formal facets of the data could be compared and evaluated. It should be clear that such a relatively simple quantitative analysis can only provide some informative corner-stones in the sphere of newspaper digitisation. A quantitative analysis can obviously not fully substitute a qualitative one.

[19] In Germany the German Digital Library (Deutsche Digitale Bibliothek, https://www.deutsche-digitale-bibliothek.de/?lang=en) is intended to act as a national aggregation service for historical newspapers.