A Bit of This, a Bit of That: Building a Genre and Topic Annotated Dataset of Historical Newspaper Articles with Soft Labels and Confidence Scores

Karin Stahel, Irenie How, Lauren Millar, Luis Paterson, Daniel Steel, Kaspar Middendorf
UC Arts Digital Lab, University of Canterbury
{karin.stahel, irenie.how, lauren.millar}@pg.canterbury.ac.nz,
{luis.paterson.nz, danjsteel}@gmail.com, kaspar.middendorf@canterbury.ac.nz

Abstract

Digitised historical newspaper collections are becoming increasingly accessible, yet their scale and diverse content still present challenges for researchers interested in specific article types or topics. In a step towards developing models to address these challenges, we have created a dataset of articles from New Zealand's *Papers Past* open data annotated with multiple genre and topic labels and annotator confidence scores. Our annotation framework aligns with the perspectivist approach to machine learning, acknowledging the subjective nature of the task and embracing the hybridity and uncertainty of genres. In this paper, we describe our sampling and annotation methods and the resulting dataset of 7,036 articles from 106 New Zealand newspapers spanning the period 1839-1903. This dataset will be used to develop interpretable classification models that enable fine-grained exploration and discovery of articles in *Papers Past* newspapers based on common aspects of form, function, and topic. The complete dataset, including un-aggregated annotations and supporting documentation, will eventually be openly released to facilitate further research.

1 Introduction

Just over 100 years ago, in an article titled "The Natural History of the Newspaper", Robert Park wrote, "The newspaper, like the modern city, is not wholly a rational product." (Park, 1923, 273). He was describing the development of the press from newsletter to political and commercial institution, but his sense of an evolving organism, something familiar yet difficult to specifically define, applies as much to the types of articles inside as it does to the newspaper as a whole. These types, or categories, of newspaper articles can be referred to as *genres*, although an agreed upon definition of this term is as difficult to pin down as the genres themselves (Chandler, 1997; Ljung, 2000; Lee, 2001; Liddle, 2015; Underwood, 2019). For our purposes in constructing a dataset of digitised historical newspaper articles annotated with genre and topic labels, we consider *genre* to be the *type of document* (or newspaper article in this case) and *topic* to be *what the document is about* (Ruthven and Pennington, 2018). We do not attempt to establish a definitive list of the types of articles found in historical newspapers but instead aim to identify articles that share characteristics of form and function along with the topics within those articles, such as "football" or "politics". We call the article categories *genres* and label them with common terms such as "editorial", "letter", or "review".

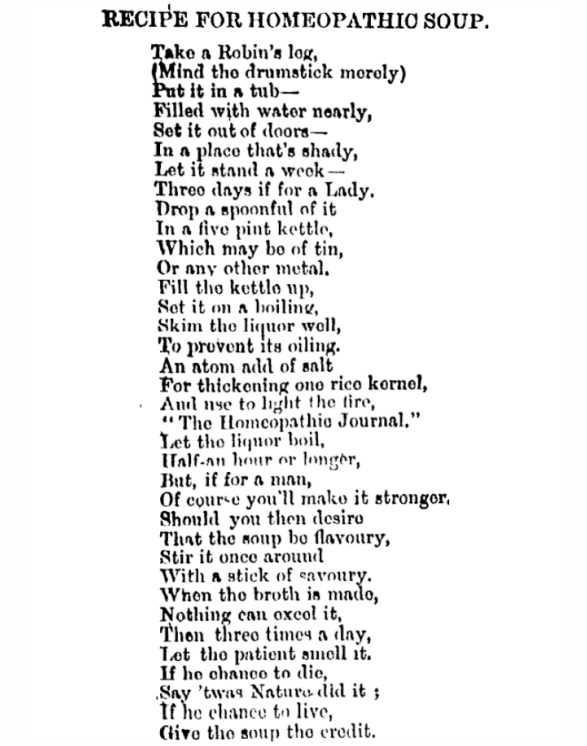
Today, we have desktop access to millions of digitised newspapers and researchers are investing significant effort in developing datasets, designing interfaces, and training state-of-the-art models to enhance these collections and extract new insights (for example Bunout et al., 2023; Dell et al., 2023; Doucet et al., 2020; Düring et al., 2024; Ehrmann et al., 2020; Lee et al., 2020). The text of many digitised newspaper collections has been made searchable through the use of Optical Character Recognition (OCR) software which, along with related technologies and post-OCR correction methods, has improved significantly in recent years (Chen and Ströbel, 2024; Reul et al., 2024; Kim et al., 2025). However, the fact remains that many historical newspaper collections contain a significant number of errors which affect the reliability of keyword search and have implications for researchers in terms of source criticism, reproducibility, and claims of representativeness (Burchardt, 2023; Cordell, 2017; Hiltunen, 2024; Hitchcock, 2013). OCR errors, misspellings, and diachronic language change also present challenges when it comes to the robust application of new methods such as Retrieval Augmented Generation (RAG) to historical text collections (Piryani et al., 2024; Thorne et al., 2024; Tran et al., 2024).

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 377–392
May 3-4, 2025

Several studies have explored the use of features such as page layout data, text statistics, TF-IDF metrics, and parts-of-speech frequencies, sometimes in addition to bag-of-words or word embedding representations, to classify news articles by genre (for example Bilgin et al., 2018; Kilner and Fitch, 2017; Langlais, 2022; Petrenz and Webber, 2011). These approaches offer alternative ways to identify and retrieve different types of texts, and work at a level of abstraction from the individual characters, which can provide resilience to OCR errors and unusual word forms. When used with interpretable machine learning methods and in transparent pipelines they can provide valuable insights into the characteristics of different genres and the ways in which these genres change and evolve (Bilgin et al., 2018; Broersma and Harbers, 2018). Misclassifications can also be informative by revealing where and how genres overlap or surfacing unusual examples that would otherwise be difficult to discover (Bamman et al., 2024; Blankenship, 2024; Langlais, 2022).

The subjectivity and hybridity of genre and its fluidity across time make genre classification a challenging and interesting problem (Blankenship and Cordell, 2024; Crowston and Kwasnik, 2004; Langlais, 2022; Underwood et al., 2013). The example shown in Figure 1, a humorous poem that contains aspects of a recipe and advice, illustrates these challenges. This text might be labeled in different ways depending on the perspective of the annotator, and classification using traditional supervised machine learning methods trained on a single hard label per article could lose valuable information. Soft labels and confidence scores from multiple annotators provide a way to better reflect the human perspective of genre by capturing its subjectivity and hybridity (Collins et al., 2022; de Vries and Thierens, 2024). This approach follows the perspectivist paradigm described by Cabitza et al. (2023), which builds on previous work encouraging consideration of human label variation in machine learning model training and evaluation (Aroyo and Welty, 2015; Basile, 2020; Fornaciari et al., 2021; Plank, 2022; Uma et al., 2021).[1] In making room for disagreement, perspective, and subjectivity, such an approach has been argued to improve model calibration, representation, and evaluation (Basile et al., 2021; Fleisig et al., 2024).

In this paper, we describe the construction and



RECIPE FOR HOMEOPATHIC SOUP.

Take a Robin's leg,
(Mind the drumstick merely)
Put it in a tub—
Filled with water nearly,
Set it out of doors—
In a place that's shady,
Let it stand a week—
Three days if for a Lady.
Drop a spoonful of it
In a five pint kettle,
Which may be of tin,
Or any other metal.
Fill the kettle up,
Set it on a boiling,
Skim the liquor well,
To prevent its oiling.
An atom add of salt
For thickening one rice kernel,
And use to light the fire,
"The Homeopathic Journal."
Let the liquor boil,
Half-an hour or longer,
But, if for a man,
Of course you'll make it stronger.
Should you then desire
That the soup be flavoury,
Stir it once around
With a stick of savoury.
When the broth is made,
Nothing can excel it,
Then three times a day,
Let the patient smell it.
If he chance to die,
Say 'twas Nature did it;
If he chance to live,
Give the soup the credit.

Figure 1: A humorous poem that could also be labeled as a recipe and advice. *North Otago Times*, Volume XXVI, Issue 1877, 2 May 1878, Page 4.

features of a genre and topic annotated dataset of digitised historical newspaper articles sampled from the National Library of New Zealand's *Papers Past* open data (National Library of New Zealand Te Puna Mātauranga o Aotearoa, 2024). The resulting dataset includes soft genre labels, annotator confidence scores, and topic labels and covers 106 New Zealand newspaper titles from the period 1839-1903. The final dataset will be made publicly available and, as far as we are aware, will be the first openly released dataset of digitised historical newspaper articles annotated in this way. It is a key part of an ongoing project focused on developing interpretable genre classification models to enhance the discovery and analysis of articles in *Papers Past* newspapers.

The key contributions of this work are: (1) a large dataset of more than 7,000 historical newspaper articles with un-aggregated soft genre labels, annotator confidence scores, and topic labels, (2) a detailed description of the sampling and annotation process and results including the genre and topic labels and inter-annotator agreement, and (3) a discussion of the challenges and limitations associated with the development of the dataset.

[1] See also: *The Perspectivist Data Manifesto.*

2 Methods

2.1 Data

Our dataset is a sample of the *Papers Past* open data, a collection of historical New Zealand newspapers released in METS/ALTO XML format by the National Library of New Zealand (National Library of New Zealand Te Puna Mātauranga o Aotearoa, 2024).[2] As at February 2025, the *Papers Past* open data includes 108 newspaper titles from the period 1839-1903. The open data was processed to extract newspaper and article titles, dates, article codes, and a list of text blocks for items with the attribute `TYPE="ARTICLE"` using an approach based on code released by Wilson Black (2023).[3] "Articles" that were not associated with any text blocks (such as the titles of illustrations) were removed, along with two newspaper titles: the *Victoria Times*, which was only published once on 15 September 1841, was handwritten (lithographed) and contained only four "articles", and *Bratska Sloga* which published four issues in May and June 1899, mainly in Croatian. Our final sampling frame contained 10,811,624 articles from 106 newspapers.

2.2 Genres

From our previous (unpublished) work on identifying newspaper article genres in *Papers Past*, we had an initial list of genre terms and an understanding of which genres appear frequently in the dataset (for example, notices and reports) and which are relatively rare (such as speeches). We supplemented this knowledge by collating an inventory of article categories used in other online historical newspaper collections, published research on newspaper genres, the International Press Telecommunications Council's (IPTC) NewsCodes controlled vocabulary,[4] and genres identified by participants in a survey of *Papers Past* users (n = 200).

In some cases, the title of the article can indicate the genre, for example, "Original Poetry", "Correspondence", or the presence of the word "Chapter" in the title of fiction. We reviewed the titles of articles associated with known genres in our previous work and examined high frequency titles of the articles in our sampling frame, and used this information to identify candidate articles for each genre in our inventory. Table 1 shows the twenty most frequent article titles after normalisation by lowercasing, removing punctuation, and making "births", "deaths", and "marriages" singular. Lists of titles considered likely to be associated with each of our final set of 22 genres were used to apply "rough" labels to almost 28% of the articles in our sampling frame, as shown in Table 2.

Normalised title	Number of articles
untitled	408,415
commercial	157,149
shipping	107,181
death	100,003
sporting	98,826
mail notices	80,520
birth	77,572
cricket	67,381
australian	59,274
telegrams	56,117
marriage	50,302
football	42,247
mail notice	40,220
local and general	37,256
shipping intelligence	36,912
correspondence	35,965
shipping telegrams	35,798
telegraphic	35,362
interprovincial	34,730
australian news	34,091

Table 1: The twenty most frequent article titles in our sampling frame of articles from the *Papers Past* open data (1839-1903) after normalisation.

2.3 Sampling

A multi-stage stratified sampling approach was implemented to extract a dataset for annotation across newspaper titles and time periods. Hierarchical quota samples were used to obtain minimum numbers of articles both within and across time periods from those identified as genre candidates and from individual newspaper titles. While quota samples can introduce bias by oversampling certain substrata (Lohr, 2021), our priority was to obtain a balanced dataset across the parameters of interest (genres, newspaper titles, and time periods) for the purpose of training and testing classification models, rather than to collect a proportionally representative sample of the population (Biber, 1993). This is similar to the approach taken by Hiltunen (2021), who aimed to create a balanced corpus across time periods and text types in the *British Library News-*

[2] See *What is METS/ALTO?* for information about this format.

[3] The *Papers Past* newspapers data currently includes basic genre tags in the form of `ARTICLE`, `ADVERTISEMENT`, or `ILLUSTRATION`, which are automatically identified during the digitisation process.

[4] https://cv.iptc.org/newscodes/genre/

Genre	Number of articles
News	1,131,170
Report	808,641
Notice	554,778
List	135,476
Editorial	94,063
Letter	56,121
Review	31,585
Advertisement	27,860
Fiction	24,097
Obituary	22,529
Opinion	21,682
Squib	19,043
Feature	16,980
Poetry or verse	9,727
Table or chart	8,448
Social column	4,504
Narrative humour	2,805
Advice	2,494
Speech	1,989
Recipe	1,746
Joke, riddle, puzzle	1,605
Narrative non-fiction	677
Total	**2,978,020**

Table 2: The number of articles in the sampling frame that were identified as candidates for each genre using article titles.

papers database.[5] Our target was 200 examples of each genre in the final dataset, based on the findings of Figueroa et al. (2012) who tested the performance of classification models with different size annotated training sets and found error rates decreased significantly for training sets between 80 and 200 instances but beyond 200 the error rates plateaued (Figueroa et al., 2012). To allow for instances in our sample where the candidate articles were not actual examples of the genre or were illegible, we set a sampling quota of 220 candidate articles per genre.

Six time periods were defined with the purpose of obtaining enough data in the early years where fewer newspapers are available and aligning with significant dates in the history of New Zealand's newspaper industry such as the introduction of the telegraph in 1861-1862, the establishment of the New Zealand Press Agency, the New Zealand Press Association, and the United Press Association in the 1870s, and the rapid growth in the transmission of press telegrams in the 1880s (Byrne, 1999; Grant, 2018; Hannis, 2008). The time periods used are: 1839-1861, 1862-1871, 1872-1881, 1882-1891, 1892-1901, and 1902-1903.

The multi-stage sampling approach to meet target sample quotas was implemented as follows:

1. Get the candidate articles for each genre and newspaper in each time period.

2. Take a random sample of **6 articles per newspaper per time period**. If fewer than 6 articles are available, take all the articles for that newspaper (there were no cases where this was necessary).

3. Using the remaining articles labeled as genre candidates, take a random sample in the time period to meet a minimum of **33 articles per genre per time period**. If fewer than the minimum are available for a genre in the time period, take all candidate examples for that genre.

4. Take more random samples to meet a minimum of **1,100 articles in total for the time period**.

5. Following the completion of steps 1-4 for each time period, check that a minimum total of **30 articles per newspaper** has been met, if not, sample more from newspapers where the condition has not been met to fulfill the quota.

6. Check if a minimum total of **220 candidate articles per genre** has been met, if not, sample more from the candidate articles to fulfill the genre quotas.

This process resulted in a sample of 7,885 articles, of which 7,791 could be matched to an article on the *Papers Past* website, a necessary requirement in order to display the article for annotation.

3 Annotation

3.1 Annotation interface

The annotation process is time-consuming and critical to the development of a quality dataset, which makes the choice of annotation tool a significant decision (Colucci Cante et al., 2024; Krušic, 2024; Neves and Ševa, 2021). For this project we required the ability to display a scrollable image of the article, support for multiple annotations per item (genres and topics), and the ability to capture confidence scores and indicate if the article was legible and if it was a single article or if it consisted of multiple items due to errors in the article segmentation process.

[5] See *British Library Newspapers*. The text types used were "Arts and entertainment", "Birth, death, marriage notices", "Business", "Classified ads', "Editorial", "News", "Sports".

After reviewing the documentation for several open source tools and considering the fit with our requirements, we ultimately decided to design our own annotation system and interface using Google Colab, with data stored in Google Cloud Storage (GCS) buckets. The interface was developed using the `ipywidgets` package and could be viewed full-screen from within Colab. This enabled flexible customisation and the cloud-based system meant the lead researcher could efficiently segment and monitor the data for each annotator using unique Google accounts. The annotation guidelines were provided as a Google Doc, which was accessible via a link on the interface. Annotations were saved to a CSV file in a GCS bucket on each click of the `Next` button in the interface. A screenshot and descriptions of key elements of the interface are provided in Appendix A.

3.2 Annotators

Six annotators (identified as Annotator 0-5) took part in the annotation process. The lead researcher, a doctoral student in data science, was Annotator 0. Three of the other annotators were recommended via word-of-mouth and two were already contacts of the lead researcher. Two of the annotators had recently completed PhDs in history, one focused on the American abolitionist movement and the other on New Zealand and the British Empire during World War I. These annotators had used digitised historical newspaper collections extensively in their research. The other annotators, a doctoral student in sociology and creative practice, a masters student in linguistics, and the manager of a humanities research lab, had either used online historical newspaper collections only for casual or personal research or had not used these types of collections at all. All are native English speakers. Four of the annotators (excluding the lead researcher and the research lab manager) were employed on research assistant contracts for 20 hours and paid at the intermediate level of our university's research assistant pay scale. The annotators ranged in age from early twenties to early fifties and three identified as women, two as men, and one as non-binary.

3.3 Annotation process

An iterative approach to dataset development, where feedback is incorporated and the process is adapted based on learnings early in the annotation process is advocated in much of the literature (for example Hutchinson et al., 2021; Alex et al., 2010; Pustejovsky and Stubbs, 2013; Monarch, 2021; Klie et al., 2024). Our annotation took place in a series of stages similar to the general process described in Krušic (2024), however, a key difference of our project was that all of the annotators, with the exception of Annotator 5, worked in the same location in blocks of four hours across five days.[6] Across this week, approximately 17 hours were dedicated to annotation and three to training, discussion, and feedback.

On the first day, the annotators were introduced to the project and there was time to read through and discuss the annotation guidelines and test the interface. The annotators all received the same set of articles to annotate on the first day. This set had been curated by the lead researcher to include examples of each of the 22 genres based on the "rough" genre labels, however, these "rough" labels were used for sampling only and were not shown to the annotators in the interface. Working together on the same set of articles fostered a shared sense of the task and the annotators were able to question and discuss the application of the guidelines to specific examples.

The first annotation requirement was to indicate if an article was legible. If it wasn't, either due to it being a poor quality scan or an illustration that had been incorrectly tagged as an article at the digitisation stage, the annotator could move immediately to the next example. For the genre labels, annotators were instructed to select a primary genre, "Genre 1", that they felt was the best fit for the article from a dropdown list, along with a corresponding confidence score in the form of a percentage, also selected from a dropdown list that incremented in ten percent intervals.

Annotators were advised that they could use up to three additional labels and associated confidence scores to indicate the mix or ambiguity of genre in an article. The confidence scores were not intended to necessarily represent the proportion of a genre in the text and, as in Collins et al. (2022), we did not require that they sum to 100. In practice, however, it sometimes felt natural to approximate the proportion of genres using the confidence scores, such as in the case of an article that was 50% a list and 50% a table or chart. On the other hand, it could be equally appropriate for some articles to be assigned

[6] Due to other commitments Annotator 5 joined the group only for the first day and part of the second day, completing annotation of all of the first day's sample independently over the course of the week.

a confidence of 100% for more than one genre, for example, if a letter to the editor was written in verse and the annotator felt that it was strongly representative of both genres. This annotation approach enabled a more natural representation of the human perspectives of the texts, which can be explored in different ways when it comes to using the information to train genre classification models. In cases where it was too difficult to identify a genre, annotators could leave the genre fields blank and complete the topic labels only, if possible. Free text fields were provided for entering topic terms and annotators were asked to use their judgement to select up to four representative words from the article text or title, or use a general topic word if more appropriate (for example, "politics" or "education"). The instructions indicated that only a single, lowercase word should be entered in each field without punctuation, although this was not enforced in the interface or emphasised during training as data cleaning and normalisation steps could be applied later.

The annotators completed between 64 and 137 articles in three hours on the first day of annotation, at an average rate of 30 articles per hour. At the start of the second day the annotation team discussed areas of disagreement and difficult cases. Minor clarifications were made to the annotation guidelines as a result, including the instruction that the label "various" could be entered to indicate articles where there were more topics than could be easily identified. The annotations from day one were retained in the dataset following review by the lead researcher. On day two, two groups of annotators each worked on a common set of articles and the results were again reviewed by the lead researcher and discussed as a team. By this point, the annotators were familiar and comfortable with the task and further feedback was minimal. The annotators completed between 102 and 140 articles in three hours on day two, at an average rate of 38 articles per hour. On subsequent days, each annotator was given a different set of articles. The lead researcher annotated every article in the sampled dataset, which took an additional 138 hours at an average rate of 52 articles per hour.

4 Annotation results

Of the 7,791 articles annotated, 652 marked as "Illegible" and 103 that were not labeled with a primary genre were removed from the dataset. The remaining dataset contained 7,036 articles annotated with a primary genre label, with 2,225 (32%) articles at least double-annotated. Table 3 shows the frequency of articles by the number of annotators.

No. annotators	No. articles	Dataset %
1	4,811	68%
2	1,876	27%
3	254	4%
4	25	<1%
5	6	<1%
6	64	1%
Total	**7,036**	**100%**

Table 3: The number and percentage of articles in the dataset by number of annotators (rounded to the nearest whole percent).

4.1 Genre annotations

Of the 2,225 articles with at least two annotators there are 1,473 articles where the primary genre selection ("Genre 1" in the annotation interface) is the same for all annotators (a percentage agreement of 66.2%). To assess the quality of the overall annotated dataset we used the Krippendorff's α inter-annotator agreement metric (Hayes and Krippendorff, 2007; Krippendorff, 2019), which is recommended for its versatility and ability to handle missing data and more than two annotations per observation (Klie et al., 2024; Marzi et al., 2024; Monarch, 2021). An α value of 1 indicates perfect agreement, 0 is agreement similar to what could be expected from random annotation, and negative values indicate systematic disagreement (Artstein, 2017; Marzi et al., 2024).[7]

Krippendorff's α scores were computed for the first and second days' annotations, and for the full annotated dataset (see Table 4). Two different approaches were used to select a single genre label for each annotator and article combination. The first approach was to simply calculate α for the genre label selected by each annotator in the "Genre 1" position, which we called the primary genre. Our second approach was more complex and involved selecting a single label for each annotator based on consensus across all annotations for the article, with a position based tie-breaker. This is similar to the "tie-breaking plurality rules" (TBP rules) found in the domain of social choice theory (Saitoh,

[7]The metric was calculated using both the *K-Alpha Calculator* developed by Marzi et al. (2024) and a Python script based on method "C" in Krippendorff (2011). The Python script was developed with assistance from Claude 3.5 Sonnet.

2022). The most common genre label across all annotators in any of the four genre positions was selected, with position only relevant for tie-breaking. In breaking ties, genres that were selected more frequently in earlier positions across all of the annotators were prioritised, as illustrated in Figure 2. If there were no shared genre labels, the annotator's primary genre selection was used. The consensus genre approach serves to normalise the effect of individual annotator preferences where the choice of primary genre can be arbitrary, for example a "Notice" that is equally an "Advertisement", or our example from Figure 1, which might reasonably be labeled with "Recipe", "Advice", or "Poetry or verse" in the primary genre position.

Annotated dataset	Primary genre α	Consensus genre α
Day 1	0.70	0.88
Day 2	0.60	0.77
Full	0.66	0.86

Table 4: Krippendorff's α scores for the annotations completed on day 1 and 2, and the full dataset, using either the primary genre or the consensus genre label for each annotator. As Annotator 0 designed the annotation scheme and conducted the training, their annotations are excluded from the day 1 and 2 metrics.

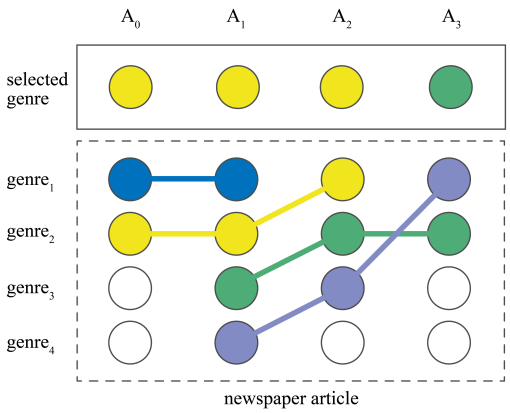


Figure 2: An illustration of the consensus approach with position based tie-breaking used to select a single genre label per annotator for each article.

Two of the α scores shown in Table 4 are slightly outside the range (0.67-0.79) considered "moderate agreement" (Marzi et al., 2024), however, we were pleased with the results given the subjective nature of the task and data. The agreement is also relatively consistent between day one and the full dataset, as is the improvement in the α score for the consensus genre compared to the primary genre. As noted by Klie et al. (2024) and Amidei et al. (2019), the interpretation of agreement results and thresholds depends on the task. We are not aware of any directly comparable datasets of genre annotated historical newspaper articles to benchmark against, however, the α scores we have achieved are higher than those reported for many other datasets of human annotated text (Klie et al., 2024). Asheghi et al. (2014) designed and evaluated a genre annotated corpus of web pages, with one of their motivations being the low agreement for existing genre-labeled corpora (Krippendorff's α scores of 0.56 and 0.55 are reported for two of the cited collections). Asheghi et al. (2014) achieved a Fleiss's kappa score (Fleiss, 1971) of 0.874 for their full dataset annotated with 15 web genres by 42 annotators.

Following Asheghi et al. (2014), we also computed per-genre agreement scores to provide insight into the genres with high and low consensus. In this case, agreement is measured with a binary approach, where the presence of the target genre in an annotator's selections for an article is coded "1" and its absence is coded "0" (or "NA" where an annotator didn't label an article). The results are shown in the first column of Table 5. Interestingly, like Asheghi et al. (2014), "Recipe" achieved the highest agreement, with α of 0.93. The lowest agreement was for "Opinion" with α of 0.36.

The accuracy of the "rough" labels used to identify genre candidates based on article titles can also be seen in Table 5. The "Genre candidates" column shows the number of candidate articles for each genre in the final annotated dataset and "Primary genre matches" counts articles where an annotated primary genre selection matched the "rough" label. "Total support primary genre" shows the number of articles where a genre is selected as the primary label with at least 90% confidence by at least one annotator. Based on this, we can see that the minimum target of 200 high quality examples was not met for several of the genres. The distribution of genre labels across the dataset will be explored further and shortfalls will be addressed with additional sampling and annotation prior to open release of the dataset.

4.2 Genre confidence scores

The distributions of annotator confidence scores for the genre labels, shown in Figure 3, are interesting to consider in relation to the per-genre agreement scores (Table 5). "Notice" has the highest possi-

Genre	α	Genre candidates (support)	Primary genre matches (support)	Genre candidate accuracy (%)	Total support primary genre $\geq$ 90% conf.
Recipe	0.93	210	169	80.48	171
Letter	0.86	218	194	88.99	453
Fiction	0.84	205	166	80.98	202
Poetry or verse	0.83	211	197	93.36	251
Table or chart	0.81	201	166	82.59	235
Advertisement	0.78	183	27	14.75	146
Obituary	0.76	214	96	44.86	102
Joke, riddle or puzzle	0.74	206	184	89.32	227
Review	0.67	206	146	70.87	193
List	0.65	185	126	68.11	367
Narrative humour	0.64	203	112	55.17	175
Squib	0.63	208	143	68.75	163
Speech	0.63	200	58	29.00	70
Editorial	0.62	178	149	83.71	386
Social column	0.61	203	173	85.22	199
Advice	0.60	197	56	28.43	114
News	0.57	361	227	62.88	689
Narrative non-fiction	0.57	168	117	69.64	183
Notice	0.55	248	195	78.63	810
Report	0.53	281	195	69.40	831
Feature	0.52	210	71	33.81	179
Opinion	0.36	200	90	45.00	286
Total		**7,036**	**3,057**	**64.73 (mean)**	**6,432**

Table 5: Metrics for each genre in the annotated dataset including per-genre Krippendorff's α, support for genre candidates identified using article titles and matches with an annotated primary genre label, along with corresponding accuracy, and total support for each genre based on primary genre selections with a confidence of 90% or greater.

ble median confidence score of 1.0, yet one of the lowest α scores (0.55). There are several possible reasons for this, including its high frequency and potential for overlap with genres such as "Advertisement", "Table or chart", and "List". The bimodal distributions and high agreement scores for "Advertisement" and "Table or chart" suggest they appear as distinctive elements but are often less representative of an article as a whole. The long tails for most genres reflect the hybridity of historical newspaper articles and the effect of inaccurate article segmentation, with the extent of each to be explored in future work. "Speech" shows a flat distribution and has the lowest median confidence score of 0.7. Empirically, a possible reason for this is the fact that speeches are often reported in the third person, making it more difficult for annotators to be confident about the selection of the genre. In addition, speeches are often not quoted in full, but form snippets of a larger context within other genres such as "Report" or "News".

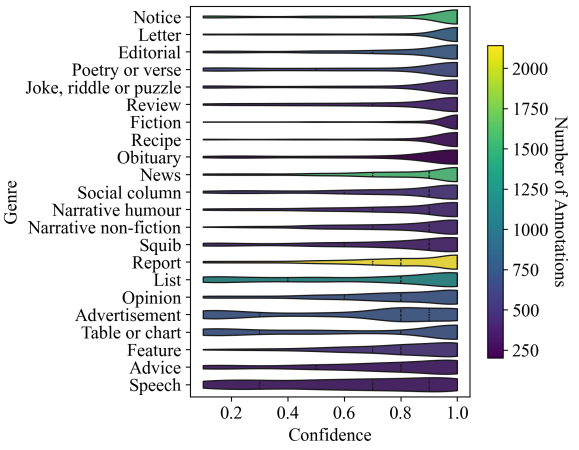


Figure 3: The distribution of annotator confidence scores and the number of annotations for each genre. The plots are sorted by median confidence score followed by the number of annotations.

4.3 Topic annotations

The free text topic annotations provide an additional perspective of the data from the same annotators, which can be useful for exploratory analysis and modeling. Although the free text format results in a large and sparse set of labels, it provides flexibility for use with a variety of methods such as topic modeling and clustering, and the terms can be mapped to a reduced set using word embeddings. Following normalisation, there are 4,583 unique topic terms from 24,687 total topic annotations in the full dataset.[8] The portion of the dataset that was

[8]Normalisation involved transliterating special characters, lowercasing, removing punctuation, concatenating multi-word annotations, and lemmatizing using WordNet.

at least double-annotated (2,225 articles, 32%) contains 2,923 unique terms, with 1,070 (37%) used by more than one annotator.

The top 20 topics according to the number of articles where two or more annotators agreed on the topic term are shown in Figure 4. The significance of the "various" label is evident, it was applied to 1,314 articles, nearly three times as many as the next most frequent term in the dataset, "politics". Figure 4 also shows the number of unique annotators who applied each topic term at least once, and the total number of articles for each term.

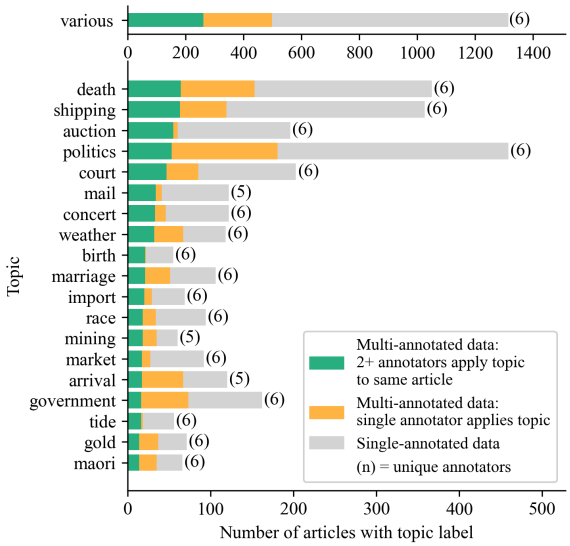


Figure 4: The top 20 topic terms (after normalisation) by the number of articles where two or more annotators agreed on the term. The numbers in parentheses show how many annotators used each topic term overall. The width of the bars shows the total number of articles where each topic term was applied.

5 Discussion

In this study, we adopted a perspectivist approach to annotating a dataset of historical newspaper articles with soft genre and topic labels and annotator confidence scores. The approach allowed us to capture the inherent subjectivity and hybridity of genre and the resulting dataset contains a wealth of information that can be used in the development and testing of genre classification models for historical newspaper texts. While we have reported Krippendorff's α using two approaches to identifying a primary genre label, there is much to be learnt from further analysis of agreement across the genre and topic labels and between annotators. The dataset lends itself to the exploration of new methods for evaluating data quality with multiple soft labels, an active area of research (Fleisig et al., 2024; Rizzi et al., 2024).

In ongoing work the annotated dataset will be used to develop interpretable classification models that enable a more fine-grained exploration of articles in *Papers Past* newspapers based on common aspects of form, function, and topic. Our focus on interpretable methods is motivated by several factors including improved transparency and reproducibility of results and the value to researchers of being able to understand and interrogate the combinations of features that contribute to an article's classification.

As emphasised by Broersma and Harbers (2018), transparent machine learning methods can support rather than replace qualitative historical research. They can be used to test hypotheses at scale and across different dimensions such as time periods, regions, or newspapers, or reveal patterns worthy of closer investigation (Broersma and Harbers, 2018). Our annotation approach enhances the exploratory potential of subsequent models by enabling more of the complexity of genre and topic to be reflected in the training data.

6 Conclusion

The sampling and annotation process described in this paper, and the resulting dataset of more than 7,000 articles from New Zealand newspapers spanning the period 1839-1903, will be of interest to researchers in digital humanities and computational linguistics, as well as those interested in exploring perspectivist approaches to machine learning. When publicly released, the final dataset will include full un-aggregated annotations and will be supported with detailed documentation that follows the recommendations of Alkemade et al. (2023), and more recently Luthra and Eskevich (2024), and includes information on the sampling and annotation process, the distribution of genres and topics, ethical considerations, metrics, and potential use cases. In ongoing work, we will use the dataset to develop interpretable classification models to further explore aspects of genre and topic in *Papers Past* newspapers.

Limitations

There are several limitations to the dataset described here, some of which will be addressed in future work and others that are inherent to the data.

The dataset is not yet balanced across genres and some newspapers are over-represented for certain genres. There are also duplicates in the dataset due to items such as advertisements being reprinted in multiple newspaper issues and titles. Some of these problems are related to the identification of genre candidates based on article titles, and these issues will be explored in ongoing work. De-duplication will be carried out using methods such as simple string matching and Jaccard similarity, with recognition that duplicates may have slightly different OCR text even when the original article is identical. Additional articles will be sampled to boost the number of the lower frequency genres such as "Speech" and "Obituary" in order to create a more balanced and versatile dataset for experiments with different subsets and combinations of genres, topics, time periods, newspaper titles, annotators, and confidence scores.

Decisions about genre and topic labels are inherently subjective and while our use of soft labels and confidence scores removed a certain pressure on annotators to select the "right" label, the annotators reported concerns about how consistent they were in their application of labels and scores. The selection of topics was particularly challenging, and annotators described having to balance spending enough time reading the text to select an appropriately representative label with the need to efficiently complete the task. A lack of context for certain genres, for example "Fiction" which might be a single chapter from a serialised work, further complicated this task. Some of the annotators said they would have found a predefined list of topics helpful, although others felt that the need to choose their own labels encouraged an engagement with the text that was also beneficial for making decisions about the genre. Related to the difficulty of the task is the issue of annotator fatigue and the potential for reduced focus and accuracy. While we tried to manage this by working in blocks of a maximum of four hours and creating a supportive and engaging environment, the annotation task required significant concentration and the annotators agreed that four hours was about the maximum that they could work effectively in a day. All of these issues could impact the quality and consistency of the annotations in the final dataset. Where we selected a single label for each annotator based on the consensus approach, we have not evaluated the impact of using confidence scores as weights or thresholds, and this is also something to be explored in future work.

As noted by Krušic (2024), working in the context of historical language increases the difficulty of annotation tasks, even when there is a level of familiarity with the sources. We were sometimes surprised by the difficulty of interpreting the intent of certain articles, for example deciding if a text was intended as serious advice or humour. We also often had difficulty finding the humour in items that we knew from other cues were obviously intended as jokes.[9] Annotators with different backgrounds and experience may have interpreted these articles differently, which is a limitation of this type of human annotated dataset.

Ethical considerations

The dataset described in this paper is sourced from digitised historical newspaper articles in the National Library of New Zealand's *Papers Past* open data collection. The articles were published in New Zealand between 1839 and 1903 and have no known copyright. The annotators were recruited as research assistants and were employed and paid for their work using established employment contracts and pay scales.

Early New Zealand newspapers predominately represent the perspective and concerns of the colonial settlers and this, in the context of the social and political conditions of the time, will be considered and documented when sharing the dataset described here. The articles in our dataset contain references to events, legislation, and attitudes that today we disagree with or consider to be outdated, harmful, or in various ways culturally sensitive. As described in this paper, the dataset will be used in our ongoing work to develop interpretable classification models that enable transparent discovery of articles in *Papers Past* newspapers. This focus on interpretability and transparency extends to the respectful treatment of culturally significant and sensitive material in the dataset. The frameworks proposed by Alkemade et al. (2023) and Luthra and Eskevich (2024) will be used to document cultural considerations and acknowledgements.

Acknowledgments

We would like to thank Dr Geoffrey Ford and the Arts Digital Lab at the University of Canterbury

[9] Nicholson (2012) provides a engaging analysis of American jokes in British nineteenth century newspapers, with many similar examples to those found in *Papers Past*.

for their support of this project and for the funding that enabled the annotators to be paid for their contribution. The advice and feedback provided by Dr Christopher Thomson, Dr James Williams, and Dr Joshua Wilson Black is also gratefully acknowledged, as is the feedback from the two anonymous reviewers. This work was further supported by the Google Cloud Research Credits program with the award GCP377961162.

References

Bea Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov. 2010. Agile Corpus Annotation in Practice: An Overview of Manual and Automatic Annotation of CVs. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 29–37. Association for Computational Linguistics.

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudecker, Giulia Osti, and Daniel Van Strien. 2023. Datasheets for Digital Cultural Heritage Datasets. *Journal of Open Humanities Data*, 9(17):1–11.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein. 2017. Inter-annotator Agreement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer Netherlands.

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2014. Designing and Evaluating a Reliable Corpus of Web Genres via Crowd-Sourcing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1339–1346. European Language Resources Association (ELRA).

David Bamman, Kent K Chang, Li Lucy, and Naitian Zhou. 2024. On Classification with Large Language Models in Cultural Analytics. In *Computational Humanities Research Conference (CHR 2024)*, pages 1–34.

Valerio Basile. 2020. It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks. In *Proceedings of the AIxIA 2020 Discussion Papers Workshop Co-Located with the the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020)*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4).

Aysenur Bilgin, Erik Tjong Kim Sang, Kim Smeenk, Laura Hollink, Jacco van Ossenbruggen, Frank Harbers, and Marcel Broersma. 2018. Utilizing a Transparency-Driven Environment Toward Trusted Automatic Genre Classification: A Case Study in Journalism History. In *2018 IEEE 14th International Conference on E-Science (e-Science)*, pages 486–496.

Avery Blankenship. 2024. What We Didn't Know a Recipe Could Be: Political Commentary, Machine Learning Models, and the Fluidity of Form in Nineteenth-Century Newspaper Recipes. *Journal of Cultural Analytics*, 9(1).

Avery Blankenship and Ryan Cordell. 2024. Word Embedding Models and the Hybridity of Newspaper Genres. *The American Historical Review*, 129(1):148–152.

Marcel Broersma and Frank Harbers. 2018. Exploring Machine Learning to Study the Long-Term Transformation of News: Digital newspaper archives, journalism history, and algorithmic transparency. *Digital Journalism*, 6(9):1150–1164.

Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors. 2023. *Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology*, volume 3 of *Studies in Digital History and Hermeneutics*. De Gruyter Oldenbourg.

Jørgen Burchardt. 2023. Are Searches in OCR-generated Archives Trustworthy?: An Analysis of Digital Newspaper Archives. *Jahrbuch für Wirtschaftsgeschichte / Economic History Yearbook*, 64(1):31–54.

Jeb Byrne. 1999. The Comparative Development of Newspapers in New Zealand and the United States in the Nineteenth Century. *American Studies International*, 37(1):55.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Daniel Chandler. 1997. An Introduction to Genre Theory.

Yung-Hsin Chen and Phillip B. Ströbel. 2024. TrOCR Meets Language Models: An End-to-End Post-correction Approach. In *Document Analysis and*

Recognition – ICDAR 2024 Workshops, pages 12–26. Springer Nature Switzerland.

Katherine M. Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and Learning with Soft Labels from Every Annotator. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 10(1):40–52.

Luigi Colucci Cante, Salvatore D'Angelo, Beniamino Di Martino, and Mariangela Graziano. 2024. Text Annotation Tools: A Comprehensive Review and Comparative Analysis. In *Complex, Intelligent and Software Intensive Systems*, pages 353–362. Springer Nature Switzerland.

Ryan Cordell. 2017. "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History*, 20(1):188–225.

Kevin Crowston and Barbara Kwasnik. 2004. A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004*, pages 1–9.

Sjoerd de Vries and Dirk Thierens. 2024. Learning with Confidence: Training Better Classifiers from Soft Labels. *Preprint*, arXiv:2409.16071.

Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers. *Preprint*, arXiv:2308.12477.

Antoine Doucet, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, Guenter Muehlberger, Eva Pfanzelter, Marie-Eve Therenty, Hannu Toivonen, and Mikko Tolonen. 2020. NewsEye: A digital investigator for historical newspapers. In *Digital Humanities 2020 (DH 2020)*. Alliance of Digital Humanities Organizations (ADHO).

Marten Düring, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity. Introducing the impresso interface for the exploration of semantically enriched historical newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 0(0):35–55.

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Ströbel, and Raphaël Barman. 2020. Language Resources for Historical Newspapers: The Impresso Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 958–968. European Language Resources Association (ELRA).

Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1):8.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels. *Preprint*, arXiv:2405.05860.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597. Association for Computational Linguistics.

Ian F. Grant. 2018. *Lasting Impressions: The Story of New Zealand's Newspapers, 1840-1920*. Fraser Books.

Grant Hannis. 2008. The New Zealand Press Association 1880–2006: The Rise and Fall of a Co-operative Model for News Gathering. *Australian Economic History Review*, 48(1):47–67.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89.

Turo Hiltunen. 2021. Exploring sub-register variation in Victorian newspapers: Evidence from the *British Library Newspapers* database. In Elena Seoane and Douglas Biber, editors, *Corpus-Based Approaches to Register Variation*, number 103 in Studies in Corpus Linguistics, pages 313–338. John Benjamins Publishing Company.

Turo Hiltunen. 2024. Early newspapers as data for corpus linguistics (and Digital Humanities): Issues in using the British Library Newspapers database as a corpus. In Mark Kaunisto and Marco Schilk, editors, *Challenges in Corpus Linguistics: Rethinking Corpus Compilation and Analysis*, volume 118 of *Studies in Corpus Linguistics*, pages 68–88. John Benjamins Publishing Company.

Tim Hitchcock. 2013. Confronting the Digital. *Cultural and Social History*, 10(1):9–23.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 560–575. Association for Computing Machinery.

Kerry Kilner and Kent Fitch. 2017. Searching for My Lady's Bonnet: Discovering poetry in the National Library of Australia's newspapers database. *Digital Scholarship in the Humanities*, 32:i69–i83.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. *Preprint*, arXiv:2501.11623.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing Dataset Annotation Quality Management in the Wild. *Computational Linguistics*, 50(3):817–866.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition. SAGE Publications, Inc.

Lucija Krušic. 2024. Constructing a Sentiment-Annotated Corpus of Austrian Historical Newspapers: Challenges, Tools, and Annotator Experience. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 51–62. Association for Computational Linguistics.

Pierre-Carl Langlais. 2022. Classified News: Revisiting the history of newspaper genre with supervised models. In *Digitised Newspapers - A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology*, pages 195–226. De Gruyter.

Benjamin Lee, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. 2020. The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. *Preprint*, arXiv:2005.01583.

David YW Lee. 2001. Genres, Registers, Text Types, Domain and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology*, 5(3):37–72.

Dallas Liddle. 2015. Genre: "Distant Reading" and the Goals of Periodicals Research. *Victorian Periodicals Review*, 48(3):383–402.

Magnus Ljung. 2000. Newspaper Genres and Newspaper English. In Friedrich Ungerer, editor, *English Media Texts – Past and Present: Language and Textual Structure*, Pragmatics & Beyond New Series, pages 131–150. John Benjamins Publishing Company.

Sharon L. Lohr. 2021. *Sampling: Design and Analysis*, third edition. Chapman and Hall/CRC.

Mrinalini Luthra and Maria Eskevich. 2024. Data-Envelopes for Cultural Heritage: Going beyond Datasheets. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 52–65. ELRA and ICCL.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Robert (Munro) Monarch. 2021. *Human-In-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Manning Publications Co. LLC.

National Library of New Zealand Te Puna Mātauranga o Aotearoa. 2024. Papers Past newspaper open data.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1):146–163.

Bob Nicholson. 2012. Jonathan's Jokes: American humour in the late-Victorian press. *Media History*, 18(1):33–49.

Robert E. Park. 1923. The Natural History of the Newspaper. *American Journal of Sociology*, 29(3):273–289.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):387–393.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. ChroniclingAmericaQA: A Large-scale Question Answering Dataset based on Historical American Newspaper Pages. *Preprint*, arXiv:2403.17859.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning*, first edition: third release edition. O'Reilly Media, Inc.

Christian Reul, Maximilian Nöth, Herbert Baier, Kevin Chadbourne, and Florian Langhanki. 2024. Human-Centred Open-Source Automatic Text Recognition for the Humanities with OCR4all. In *Proceedings of the Workshop on Humanities-Centred Artificial Intelligence (CHAI 2024)*.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: An assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94. ELRA and ICCL.

Ian Ruthven and Diane Pennington. 2018. Information attributes. In Katriina Byström, Jannica Heinström, and Ian Ruthven, editors, *Information at Work: Information Management in the Workplace*. Facet Publishing.

Hiroki Saitoh. 2022. Characterization of tie-breaking plurality rules. *Social Choice and Welfare*, 59(1):139–173.

William Thorne, Ambrose Robinson, Bohua Peng, Chenghua Lin, and Diana Maynard. 2024. Increasing the Difficulty of Automatically Generated Questions via Reinforcement Learning with Synthetic Preference for Cost-Effective Cultural Heritage Dataset Generation. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 450–462. Association for Computational Linguistics.

The Trung Tran, Carlos-Emiliano González-Gallardo, and Antoine Doucet. 2024. Retrieval Augmented Generation for Historical Newspapers. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Ted Underwood, Michael L. Black, Loretta Auvil, and Boris Capitanu. 2013. Mapping mutable genres in structurally complex volumes. In *2013 IEEE International Conference on Big Data*, pages 95–103.

Joshua Wilson Black. 2023. Creating specialized corpora from digitized historical newspaper archives: An iterative bootstrapping approach. *Digital Scholarship in the Humanities*, 38(2):779–797.

A Annotation Interface

Key elements of the annotation interface and summary instructions for annotators are shown in Figure 5 and described below.

1. **Previous button**: Go back to a previously annotated article to review it or make changes.

2. **Document links**: In the left-hand panel of the interface you will find a link to the annotation guidelines and to a Google Doc that you can use as a scratchpad to take notes during the annotation process. For example, you might want to record the details of an article that was difficult to categorise or note an interesting example of a multi-genre article such as a letter to the editor written in verse.

3. **Article details**: This section shows the article title, the newspaper title, and a code consisting of the newspaper title, date, and article number. There is also a link to view the article on the Papers Past website, although this shouldn't be necessary.

4. **Article image**: This area shows the scanned image of the original article and can be scrolled both vertically and horizontally.

5. **Next button**: Once you have completed the annotation fields (6, 7, 8, 9), click "Next" to save the annotations and move to the next article.

6. **Is the article legible?**: Some articles may be difficult or impossible to read due to the quality of the scan, or they may be illustrations or photographs that have been mislabeled at the digitisation stage. If the article is illegible, indicate the reason and click "Next" to move to the next example.

7. **Single article or multiple items?**: Sometimes an "article" in Papers Past actually consists of multiple unrelated items that haven't been separated during the page segmentation process. The decision here is: is this a single item or a column of items that are distinct examples of a single primary genre (for example, a squib, letters, or news) OR are they obviously different and distinct items of different genres? If they are obviously different items/articles please select "Multiple" here. Single articles that are a hybrid of genres should be marked as "Single" with the hybridity or uncertainty indicated using multiple genre labels and associated confidence scores.

8. **Genre labels**: Select one of the 22 genres listed in the "Genre 1" dropdown as the primary genre and use the "Confidence" score to indicate your confidence in the fit of this genre label. In some cases, more than one genre might be applicable to the text. In these situations, use the additional genre labels and confidence scores to indicate the mix or ambiguity of genre for the text. The confidence scores do not necessarily represent the proportion of the genre in the text and do not need to sum to 100. If it's not possible to identify a genre for the article, you can leave these fields blank and complete the topic labels only.

9. **Topics**: Select up to four topic words that best represent the most obvious topics in the article. Use your judgement to select representative words from the article text or title, or use a more general topic word if appropriate (for example, "politics" or "education"). Try to enter only a single word in each box. If the article contains too many topics to easily identify, you can enter "various" as one of the words to indicate this. For the example shown, the topic words might be Topic 1: commercial, Topic 2: cattle, Topic 3: flour, Topic 4: retail. Enter the words in lowercase and without punctuation.

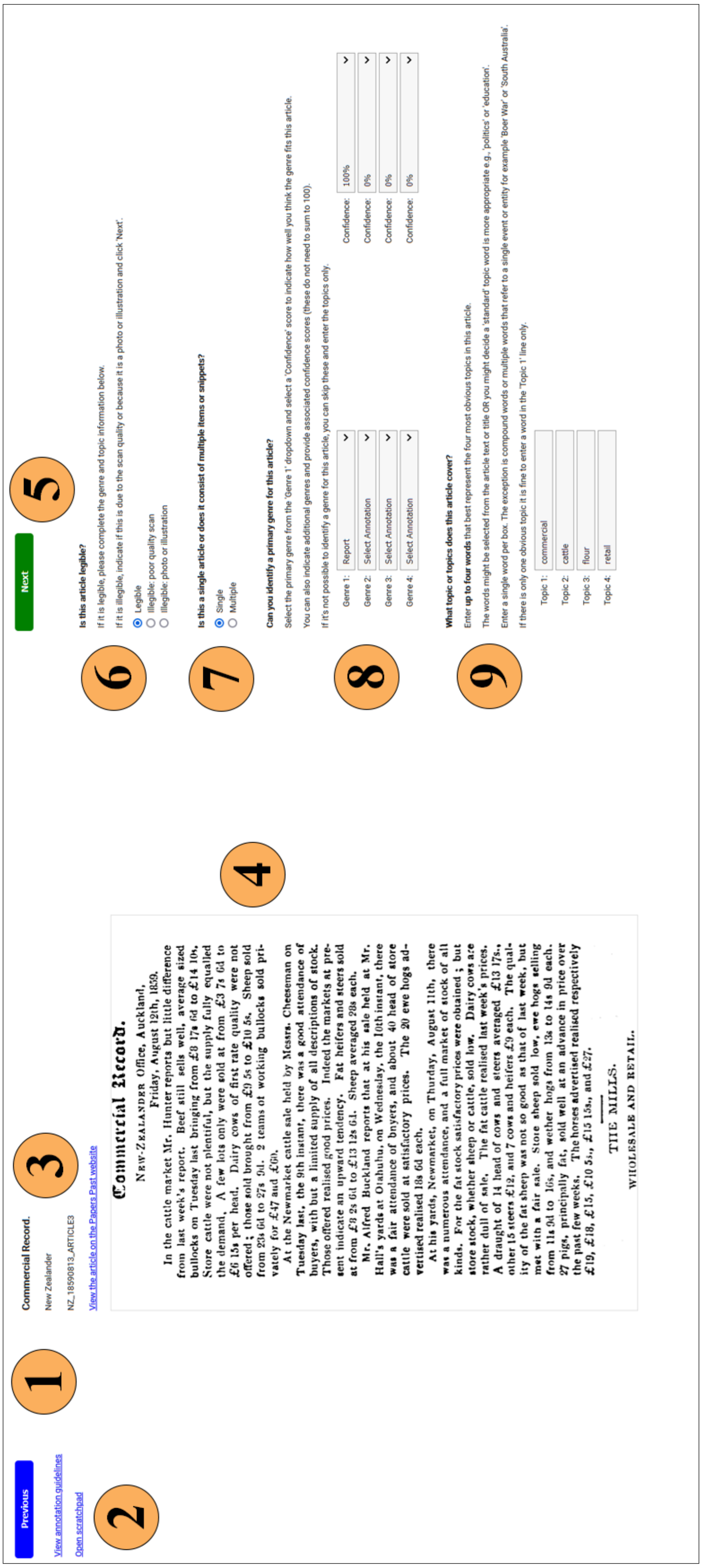


Figure 5: A screenshot of the annotation interface implemented using Google Colab.