

# A Humorous View into the Past: The Old Jokes Archive

Mark M Hall<sup>1</sup>[0000-0003-0081-4277] and Bob Nicholson<sup>2</sup>[0000-0002-0863-963X]

Department of Computer Science, Martin-Luther-University Halle-Wittenberg,  
Germany [mark.hall@informatik.uni-halle.de](mailto:mark.hall@informatik.uni-halle.de)

Department of English and History, Edge Hill University, United Kingdom  
[bob.nicholson@edgehill.ac.uk](mailto:bob.nicholson@edgehill.ac.uk)

**Abstract.** Jokes represent one of the most understudied sources about nineteenth century society. Due to their ephemeral nature they slipped from attention as soon as they were no longer funny or topical. Digitisation of newspapers and books has made them available again, but due to their short nature they are not easily accessible through current generic keyword-based newspaper search systems. In this paper we present the Old Jokes Archive, which aims to provide a digital archive focused solely on jokes. The archive will support the full process from initial text acquisition to search and finally re-use by both academic and general public users.

## 1 Introduction

Jokes represent one of the most ephemeral spoken (or written) interactions [3]. They might provoke brief laughter before the conversation moves on. They might, if they are particularly funny, even be re-told to friends, family, or colleagues. But these exchanges typically go unrecorded. While more substantial works of art and literature are carefully preserved for posterity by libraries and museums, even the most rib-tickling gags are usually disposed of and forgotten when they lose their capacity to provoke a laugh.

While jokes are often treated in a disposable fashion, they have nevertheless played important roles in many historical cultures. In nineteenth-century Britain, for instance, the possession of a good new joke represented significant cultural capital, for to be a true wit was a position of social distinction [3]. This appetite for humour was fed by the popular press and, by the 1880s, most of the best-selling newspapers and magazines in Britain featured a regular column of jokes, puns, and comic stories.

The availability of jokes in newspapers also means that, unlike longer, humorous novels and stories, jokes reached a much wider audience spanning all social classes. Thus a single joke could in one week reach as many readers as a best-selling novel might throughout its author's life-time (see for example [5] for circulation numbers on Mark Twain's work). This means that joke-based humour potentially represents a more democratic and representative view of nineteenth-century society's tastes.

At the same time, while the second half of the nineteenth century saw the introduction of copyright laws [4], this had little effect on the re-use of jokes across newspapers. Editors continued to crib jokes from other newspapers, sometimes verbatim, sometimes adapting the joke’s setting or context to suit local tastes. Due to this jokes can serve as an illustration of what and how ideas spread via the nineteenth century equivalent of viral memes [2].

The jokes themselves are a potentially invaluable source of information about historical cultures, as they are typically built upon an assumption of shared knowledge; a belief that an audience will immediately understand how a stock character (such as a mother-in-law) behaves, how a familiar social situation should ordinarily play out, or what to respond with when somebody says ‘knock-knock.’ Historians can reverse-engineer these jokes in order to uncover the ideas and attitudes that joke-writers and their editors assumed were widely held at the time. The subjects of jokes, and the dynamics of laughter (who is laughing at/with who), can also reveal valuable new insights into the power relations at work in historical communities.

Even though they represent such a rich data-source, jokes are not widely used by most historians. Many Victorianists, for instance, rarely venture beyond the cartoons of *Punch* magazine when attempting to make sense of the period’s comic culture. Millions of historical jokes have never been examined by historians and are therefore ripe for further exploration.

The chief obstacle to this research centres on the difficulty of finding and accessing historical jokes. Many were never recorded and have now been lost to us, while even those that were written down and preserved in books and newspapers are tricky to uncover. The large-scale digitisation of newspapers, books, and other print archives presents new opportunities for solving this problem. However, even with the help of keyword searches it remains difficult for researchers to locate specific jokes pertaining to their interests. For example, a keyword search for the word ‘lawyer’ in a typical digital newspaper archive will return a jumble of millions of news stories, adverts, editorials, letters, serialised stories, and poetry, within which we might also find jokes about the legal profession. At present there is no straightforward way to search specifically for historical jokes.

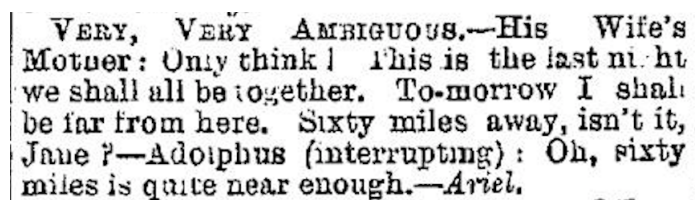
## 2 The Old Jokes Archive

The Old Jokes Archive (OJA) aims to address this gap, by providing the first, large-scale digital repository of historical humour, targeted both at academic researchers and the wider public. To support this the OJA provides a range of functionalities centred around two areas: the acquisition of the joke data and search and (re-)use functionality.

### 2.1 Acquisition

The joke acquisition starts based on the existing digitisation of newspapers and joke books, with the final goal being annotated text versions that can then be

used by researchers and the general public. Throughout the process a semi-automatic approach is used, combining automatic image and text processing with manual validation and post-processing using a crowdsourcing approach. In the OJA's precursor – the Victorian Jokes Archive – we have tested a number of crowdsourcing methods, in particular to improve the accuracy of crowdsourcing results, the experience of which will be integrated into the OJA. The joke acquisition process follows these five steps:



VERY, VERY AMBIGUOUS.--His Wife's  
Motuer: Only think! This is the last night  
we shall all be together. To-morrow I shall  
be far from here. Sixty miles away, isn't it,  
Jane?--Adolphus (interrupting): Oh, sixty  
miles is quite near enough.--Ariel.

**Fig. 1.** Example joke taken from Jokes of the Day, Lloyd's Weekly Newspaper (12/04/1891). Demonstrates the quality level common in newspaper scans.

1. **Identification** As the digitisation has focused on whole newspaper pages, the first step is in the identification of those areas of the scanned pages that contain jokes and then the splitting of those areas into the individual jokes. We are investigating automated methods for this, but have initially adapted techniques from the Digital Playbills project.
2. **Transcription** OCR will be used to create an initial transcription. Newspaper paper and typeface can be relatively poor (see Figure 1), resulting in comparatively high error rates. To deal with this we have developed error classification heuristics, that allow us to determine the quality of the OCR output. Based on this crowdsourcing users can be offered a choice to work on a transcription that has a low error rate – mostly typo-correction – or a high rate – essentially transcription from scratch. Using this both users who have significant time to invest or those who just want to do something quick and simple can be offered appropriate crowdsourcing jobs.
3. **Classification** The resulting corrected transcription is then classified using automated heuristics developed previously. The classification works at a high level, distinguishing categories such as question-and-answer, dialog, or puns. The classification is then used by the following steps to apply category-specific heuristics.
4. **Segmentation** The joke's text is segmented into chunks. The exact chunks depend on the joke category determined in the previous step, but for example for question-and-answer jokes this would be segmenting the question and the answer element, while for dialog jokes it includes identifying speakers, spoken text, and asides. Also more generally we have developed heuristics to identify joke titles and attribution.

5. **Annotation** The chunks are then, where appropriate, annotated with specific meta-data extracted from the chunk. Among others we have developed heuristics to identify dialogue speaker gender and are currently working to automatically identify social class, age, and jobs of speakers.

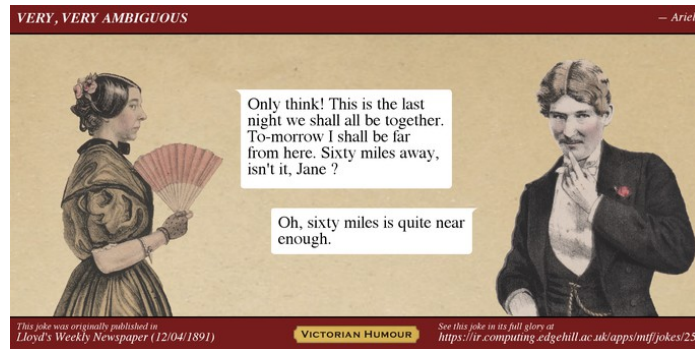
After running through all five steps, the joke texts will be publicly available through the online archive. Jokes that have been OCRed, but lack the error correction and further processing will also be available, if the user also wishes to see those.

## 2.2 Search and Use

In order to make the resulting archive available, the OJA will be available online and provide a range of access points to explore, search, and use the jokes:

- **Search** The OJA will provide a state-of-the-art faceted search system enabling users to narrow their search for jokes via keywords and via the categories and annotations created in the acquisition stage.
- **Exploration** While search works well for users who know what they want, for the general public a more open-ended, browsing based interface will be developed. Based on previous work [1] we will be developing a virtual museum of jokes through which the user can explore the available jokes.
- **Related Jokes** As described above, jokes were frequently copied from one publication to another, often with minor modifications. The OJA will provide the necessary tools to trace such copies across multiple publications and to visualise joke distribution networks.
- **Export** The OJA will provide export functionality at all points in the system, whether they be search results, exploration pages, or individual jokes. For interoperability reasons a joke-specific TEI schema will be used. Additionally the OJA will provide a workspace allowing users to save jokes to their own work area and then export that.
- **Re-interpretation** The vast majority of historic jokes tend not to have aged well in the way they are presented. The OJA will provide space for users to re-interpret the joke's text. As part of this we have also developed algorithms to automatically convert jokes into a single-panel comic image. In both cases the aim is to have these shared via social media, in part to increase the project's visibility, but also to investigate overlap and differences between popular jokes in the nineteenth century and now.

While initially the focus will be on English-language jokes, the project will be built with multi-lingual content in mind. Additionally content will be available through a permissive open-culture license to encourage use and re-use both academically and for private reasons.



**Fig. 2.** Example single-panel comic image generated automatically based on the annotated joke text derived from the joke in Figure 1. The character images are selected to match the identified gender of the speakers (mother-in-law is a woman, Adolphus is a male name). Also demonstrates the identification of the joke title and attribution in the title of the comic image.

### 3 Conclusion

Jokes represent a so far largely untapped resource for investigating a range of historical questions, from language use to the spread of ideas. The Old Jokes Archive will act as a central data-source point, starting with nineteenth century English-language jokes, but with an aim to expanding both temporally and linguistically. The tools, methods, and practices developed in the course of this project will be applicable to future archival projects that aim to 'remix' existing digitised content by extracting, organising, and re-presenting it in new ways.

While the initial focus will be on the joke's texts, the long-term aim is to also consider visual aspects in the source data, such as when jokes were accompanied by drawings.

### References

1. Hall, M.M.: Digital museum map. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J.C. (eds.) *Digital Libraries for Open Knowledge*. pp. 304–307. Springer International Publishing, Cham (2018). [https://doi.org/10.1007/978-3-030-00066-0\\_28](https://doi.org/10.1007/978-3-030-00066-0_28)
2. Nicholson, B.: 'you kick the bucket; we do the rest!': Jokes and the culture of reprinting in the transatlantic press. *Journal of Victorian Culture* **17**(3), 273–286 (2012). <https://doi.org/10.1080/13555502.2012.702664>, <http://dx.doi.org/10.1080/13555502.2012.702664>
3. Nicholson, B.: *Capital Company - Writing and Telling Jokes in Victorian Britain*. forthcoming (2019)
4. Seville, C.: *Literary Copyright Reform in Early Victorian England: The Framing of the 1842 Copyright Act*. Caombridge University Press (1999)
5. Stone, A.E.: Review: Mark twain in england, by dennis welland. *Nineteenth-Century Fiction* **34**(9), 357–359 (1979)