Building *The Old Joke Archive*

Book Section

Building The Old Joke Archive

Bob Nicholson and Mark Hall

In 1892, the writer Sarah Butler Wister went before the American people and made a heartfelt 'Plea for Seriousness'.[1] In an article for the *Atlantic Monthly,* she complained that her countrymen had developed an unhealthy obsession with humour. The greatest vice of the age, she argued, was a 'constant craving for the ludicrous' and an incessant desire to make 'a joke of everything'.[2] Worst of all, it was the 'dismal jocosity' of the period's published humour that troubled her most:

> Does the column of newspaper facetiae add to the average of daily cheerfulness? Do the funny books on railway stalls lift the burden and heat of the day, or warm the cockles of the heart against its chill? If people take comfort in exchanging such pleasantries among themselves, well and good, but to see them in print recalls Macaulay's outburst – "A wise man might talk folly like this by his fireside, but that any human being, having made such a joke, should write it down, copy it out, transmit it to the printer, correct the proof, and send it forth to the world, is enough to make us ashamed of our species".[3]

At the heart of Wister's objections to printed humour was her belief that 'nothing is more volatile and evanescent than the essence of a joke'.[4] A good jest might burst forth naturally in conversation, but its spirit could evaporate just as quickly. While more serious works of art and literature could be passed from one generation to the

next, comedy, for Wister, was condemned to decay; even the best jokes were destined to become 'weary, stale, flat, and unprofitable for merriment'.[5] The humours of the past, she concluded, were destined to be forgotten.

I shudder to think what Wister would make of *The Old Joke Archive* [*TOJA*]— a project that aims to recover, catalogue, and explore the very same jests that she so vociferously denounced. *TOJA* is an open access digital archive of historical jokes, developed in partnership between a historian, a computer scientist, and the British Library Labs. Our core mission is to make historical humour more accessible to both academic researchers and members of the public alike. All of the jokes in *TOJA* are keyword searchable and have been tagged with richly descriptive metadata, thereby allowing users to identify and filter material using categories such as a joke's place of publication, date, subject, format, and the identities and relationships of characters that appear within them. For instance, a user might search for conundrums featuring milkmen that were published in Liverpool during the 1860s. Alternatively, they might request all dialogue jokes printed in newspapers featuring a husband, a wife and the word 'hat'. As the collection grows, these tags will also allow us to 'distant read' the archive and quantitively track joke-telling trends, such as the emergence and decline of particular comic subjects, regional variations in taste, or stylistic and editorial differences between comic periodicals. *TOJA* will eventually contain hundreds of thousands of jokes from a wide range of historical sources, periods, and locations. However, for now, the pilot phase of the project focuses on recovering English-language, textual humour from the nineteenth century. Our research into the history of joke-telling in this period is still ongoing and our preliminary findings form the basis of other publications.[6] This chapter examines the construction of *The Old Joke Archive* and reflects on the methodological challenges involved in building a

repository of historical humour.

While Wister would doubtlessly disapprove of our endeavours, her ‘Plea for Seriousness’ neatly highlights the key challenges and opportunities faced by our project. Firstly, her frustrations speak to the pervasiveness of humour, as well as its social, cultural, and political significance. A society that seeks to ‘make a joke of everything’ offers rich pickings for a collector of jests, but there are challenges involved in systematically documenting a practice that was apparently so ubiquitous and diffuse. Secondly, and rather more ominously for an archival project, Wister rightly highlights the transient nature of joking. After all, most of the folly talked by our firesides is never transmitted to a printer, and even publication is no guarantee of a joke’s long-term survival. All of which leads us to the question at the heart of this chapter: how do we build an archive devoted to a historical practice that was apparently so pervasive, *and yet* so ephemeral? We begin by reflecting on the process of finding new sources of old jokes and by outlining the main forms in which they have — and have not — survived. Part two of the chapter considers the curation of jokes and the challenges involved in determining and cataloguing their subjects and genres.

Finding Jokes

The word ‘joke’ is a nebulous term that encompasses a wide variety of comic practices, texts, and expressions. In its broadest sense — a thing intended to cause amusement or laughter — we might use it to describe a prank, a cartoon, a funny passage in a novel, a line of stand-up comedy, a comic detail in a painting, a conversational quip, a piece of graffiti, and countless other humorous expressions. In

the fullness of time, *The Old Joke Archive* may expand in order include several of these comic forms. For now, however, we have adopted a narrower set of criteria that define a joke in the following terms: a short, self-contained, comic text or expression that typically terminates in a punchline. This definition is best exemplified by the short gags published in joke books and comic papers, but our criteria are expansive and flexible enough to cover other modes of publication, as well as second-hand evidence of jokes which were performed on stage or shared with friends. Our stipulation that jokes must be self-contained excludes those that were contextually dependent on their place within a wider text or performance; a funny quip, or call-back, in the middle of a narrative stand-up routine would probably not meet our criteria, but a one-liner told by the same comedian might. Our guidelines also exclude comic short stories, although the line between these texts and some of the longer jokes in our archive is ambiguous and rests on whether the story builds to a punchline. We have not implemented a strict word-count for our shortness criteria, chiefly because some gags derive their humour from a tortuously extended set-up. Similarly, while most jokes in our archive finish on a clear punchline, there are some exceptions to this rule. Victorian joke writers often explained pun-based punchlines using a parenthetical postscript, while accounts of the jests told by famous wits sometimes describe how these immortal quips were received by their audiences. Our punchline criteria also leave room for the anti-jokes and shaggy dog stories that comically subvert joke-telling conventions by ending anticlimactically. Finally, our focus on textual jokes (or spoken jokes that can be rendered, or described, textually) also excludes cartoons, caricatures, comic strips, and slapstick routines. While visual humour is a fascinating and important element of many historical cultures of joke-telling, a different archival infrastructure and interface is needed to curate and

display this material.

Historical jokes, as we define them, have survived in an eclectic range of places and formats. The jokes in *TOJA* come from historical sources that fall into one of four main categories, each of which presents us with different curatorial and technical challenges. In each case, our goal is first to *find* a prospective source of old jokes, then to *extract* the individual jokes contained within it, and finally to *annotate* each of these jokes with descriptive metadata. The end result is a record in *TOJA* for each instance of a joke, featuring: (1) an image of the joke as it appeared in its original publication; (2) a digital transcription of the joke's text in order to make it keyword searchable; and (3) a range of metadata categories which are discussed later in this chapter. While this might seem like a relatively straightforward workflow with a consistent output, the haphazard ways in which jokes have been preserved means that the tasks and tools involved in this process vary considerably depending on the nature of documents involved.

Joke Books

Existing research on the history of jokes typically concentrates on joke-focused publications, the most obvious examples being historical jest books.[7] The great advantage of these sources is that they contain a large quantity of jokes, the vast majority of which fit the criteria for our archive. The most substantial books in our collection contain thousands of jests, while even their smallest counterparts usually have dozens of items that are eligible for inclusion in *TOJA*. In many cases, these publications also have titles and bibliographic metadata that make them discoverable using library and archival catalogues. The British Library catalogue, for instance,

immediately returns 203 results for nineteenth-century publications featuring the words ‘jokes’ or ‘jests’ in their title, almost all of which are joke books. Follow-up searches for terms like ‘wit’, ‘humour’, ‘comic’, ‘laugh’, and ‘fun’ reveal thousands more. While these basic catalogue searches have proven fruitful, it is important to stress that many historical joke books are not so easily traced. For instance, one late-nineteenth-century anthology named *How To Set The Table In A Roar!* features 256 pages packed with jokes, but does not have an obviously comic keyword in its title.[8] Moreover, even though the book was published by a company based in London and Yorkshire, it does not appear in the British Library’s catalogue or, for that matter, the catalogues of other major UK libraries. A search on WorldCat returns just one result for a copy of the book held by Cornell University in New York State. This is not an uncommon problem. While joke books were theoretically subject to the same laws of legal deposit as other publications, many seem to have slipped through the net and evaded systematic preservation. This is particularly true of titles released by minor, short-lived publishing houses and flimsy texts such as chapbooks and pamphlets. This is, perhaps, a reflection of the perceived ephemerality and disposability of jokes; it may also stem from the fact that many of these anthologies were heavily based on unattributed reprints and were not identified with a specific author or editor, placing them in an ambiguous relationship with copyright. In any case, it is likely that many historical joke books have been lost forever. Many of those that have been preserved are now scattered across the globe, buried in uncatalogued special collections, or held in private hands. Finding these texts requires lengthy detective work and a dash of good fortune. We only discovered *How To Set The Table In A Roar!* when it was listed on eBay with the word ‘jokes’ in the description. As a result, while it is unlikely that *TOJA* will ever feature *all* of the jest

books printed in a particular place or period, some of the holes in our collection will gradually be plugged as more publications are discovered.

While some historical jest books are hard to track down, others were so frequently reprinted that they present an entirely different curatorial problem. Dozens of eighteenth and nineteenth-century jest books were named *Modern Joe Miller, New Joe Miller*, or *Joe Miller Up To Date* in homage to the celebrated *Joe Miller's Jests* of 1739. Some were straightforward reprints, either of the original text or one of its earlier imitators; others borrowed selectively from previous incarnations of Joe Miller while adding their own material; and several of these self-appointed sequels shared no content with the 'original' texts whatsoever. In all cases, we aim to add the book's contents to *TOJA,* even if it is a word-for-word reprint of a text that we have already processed. There are two reasons for this apparent duplication of effort. Firstly, it is almost impossible to conclusively determine the 'original' source of most jokes; even the ones in an ostensible foundation-text such as the first *Joe Miller* were probably borrowed from other texts, both ancient and modern, or plucked from oral culture. To include only the earliest instance of each joke would give a misleading impression as to their authorship and require us to assess and cross-reference the date of each individual joke before deciding whether to include or reject it. Secondly, we consider each reprint and retelling of a joke to be a meaningful historical act with its own distinctive purpose and context. Capturing this data gives historians a better understanding of how, where, when, and why particular jokes were told. As *TOJA* grows, so too will our ability to trace the origins, evolution, and circulation of individual jests. In an earlier project, Nicholson tracked the life of a single nineteenth-century joke from its apparent origins in a New York comic periodical, through its international circulation across dozens of English-language books and newspapers,

to its eventual status as a stock jest used by British politicians.[9] This was accomplished by keyword searching multiple digital archives for a phrase in the joke's punchline and then trawling through thousands of irrelevant hits. *TOJA* will simplify this process by automatically suggesting connections between similar jokes in its collection, even if they are not word-for-word reprints of one another.

Comic Periodicals

Jest books are the most concentrated source of material for *TOJA*, but they represent only a small fraction of the total number of jokes that have survived within the historical record. Comic periodicals are, potentially, an even richer source of jokes, particularly for the Victorian Era. *Punch* magazine, founded in 1841, is the most famous of these publications and is routinely cited by historians seeking to discover what made the Victorians laugh. However, while it was certainly an influential title, it is important to stress that *Punch* was not the only humorous paper in wide circulation. Donald J. Gray's bibliography of comic periodicals lists nearly four hundred titles published in Britain between 1800-1900, and this list is unlikely to be exhaustive.[10] The majority of these papers were very short-lived, some only lasting for a single issue, but many appeared on a weekly or monthly basis for years at a time, and some endured for decades. As well as supplying a large amount of material for our archive, these long-running titles make it possible to explore how, and why, jokes changed over time, or, alternatively, to pinpoint jests that arose at a very specific historical moment. Indeed, the periodicity of comic periodicals — their weekly and monthly rhythms, their ability to respond rapidly to current events, and their expected disposability — means that they were produced and consumed in

very different ways to one-off jestbooks and carried different types of comic material as a result. A witty pun about a recent parliamentary speech by Gladstone might appear in a weekly paper like *Punch,* but the topicality of this joke would make it too short-lived for inclusion in a hardbound jestbook designed to sit on bookshelves long after the Prime Minister's speech had been forgotten.

The titles in Gray's bibliography vary enormously in style, content, and target audience: from expensive, high-brow, satirical periodicals aimed at middle-class literary gentlemen, to more accessible halfpenny weeklies whose social jokes and slapstick cartoons were more suited to a popular audience. By studying key aspects of these periodicals — their price, title, branding, circulation, adverts, tone, political position — it is possible to make tentative observations about the class, gender, age, and politics of the readers who were expected to enjoy and understand the jokes within them. While similar techniques can be used to reconstruct the implied readers of jestbooks, these texts typically feature fewer contextual clues as to their intended audience. Comic periodicals are also more effective for explore geographical variations in joke telling. Some titles enjoyed a national circulation, but others were focused on a serving, and satirising a particular town or region, such as Liverpool's long-running *Porcupine* and Edinburgh's shorter-lived *Mac Punch*. In these publications, jokes about well-known local people, places, and events were far more common. Integrating a wide range of comic periodicals into *TOJA*, therefore, is desirable not simply because they increase the overall size of our archive, but because their jokes reflect different aspects of the period's comic culture and enable us to contextualise it with greater precision.

Unfortunately, the vast majority of comic papers have yet to be digitised, and some of the rarest surviving titles are in fragile condition. While it is relatively cost-

effective and practical for us to digitise joke books, particularly those already owned by researchers on the project, our current budget does not stretch to the large-scale digitisation of hundreds of comic periodicals spread across multiple libraries and archives. Each page of a jest book typically yields between five and fifteen jokes, but periodicals contain a wide range of other comic genres that do not fit *TOJA's* current selection criteria. Satirical political commentary, comic verse, parodies, burlesques, short stories, and cartoons all fall outside our purview. A typical Victorian comic periodical, modelled on *Punch's* successful formula*,* usually featured a handful of jokes (as we define them) in each weekly issue, but scattered them throughout the paper as column-filers that need to be identified and extracted one-by-one. This is a task that Victorian sub-editors knew only too well. Every week, in newspaper offices around the country, they combed the pages of newly released comic periodicals in search of jokes. Soon after, they reprinted them in dedicated humour columns with titles such as 'Wit and Humour' and 'Pickings from the Comic Press'. Just like us, these Victorian newspaper editors needed to identify self-contained, short pieces of humour that would survive the decontextualising process of being transplanted into a new publication. Observing their well-honed technique helped to sharpen our curatorial practices and also pointed to an unexpected source of new material. Curiously, most of the jokes reprinted from *Punch* were lifted from the captions that appeared underneath the magazine's cartoons, many of which functioned surprisingly well in the absence of any visual component. This kind of copying was quicker, cheaper, and technologically simpler than reproducing magazine's woodcuts, and it also seems to have lessened the risk of legal disputes; *Punch* litigiously defended the copyright of its illustrations, but generally turned a blind eye to persistent, low-level textual plagiarism.[11] *TOJA* operates in a different legal and

technical context, but the extraction of jokes from the captions of some cartoons is also a viable option for our archive.

In practice, the most efficient way for us to gather material from comic periodicals is also inspired by the practices of nineteenth-century editors. While some publications clipped jokes directly from *Punch*, others seem to have bypassed this task and copied the selections made by other newspapers. *TOJA* adopts a similar approach and piggy-backs on the work of Victorian editors by collecting material from comic periodicals that was reprinted in daily and weekly newspapers. These jokes were often published with attributions, which allows us to trace them to an original publication. As *TOJA* grows, it will also become increasingly capable of pinpointing the sources of unattributed jokes — the work of one diligent sub-editor is sometimes enough to reveal the uncredited borrowings made by dozens of his rivals. There are clear pros and cons to this second-hand curation of the comic press. On the plus side, newspapers have been digitised more extensively than comic periodicals and are therefore much cheaper and easier for us to access and process. As Nicholson has explored elsewhere, one of the most efficient ways to build a new digital archive is to creatively 'remix' material that has already been scanned by other projects, but which is not well-served by their interfaces and metadata schema.[12] Importing newspaper joke columns also bypasses the manual task of reading through the comic press and picking out the jokes ourselves, allowing us to add material to the archive more quickly. Outsourcing our curatorial judgement to Victorian sub-editors is an unconventional strategy, but our experiments suggest that their choices closely mirror our own. Crucially, we do not rely on the selections of a single editor, but on the collective clippings made by dozens of papers. In the process, we also gain valuable new datapoints that help to reveal the true popularity,

circulation, and audience of these jokes. The unavoidable downside of this approach is that Victorian editors typically focused their attention on a narrow range of publications — *Punch* looms large, closely followed by other London-based weeklies like *Judy*, *Fun,* and *Ariel.* At present, the pilot stage of *TOJA* replicates this focus because it offers the most efficient method of harvesting jokes, but also because it captures the most widely circulated jests of the age. However, as the archive develops, we aim to address this imbalance and manually process jokes from obscurer Victorian periodicals that were left undisturbed by the scissors of the period's editors.

Newspaper Humour

Newspapers are arguably the most promising source of nineteenth-century jokes, and not just because they grant us second-hand access to the highlights of comic periodicals. More jokes have been preserved in these publications than in all comic periodicals and joke books put together. Jokes and other comic clippings appeared sporadically in newspapers from the earliest days of the press, where they functioned as useful column-fillers. However, from the 1870s onwards, organised humour columns became a staple feature of provincial and national newspapers alike. A typical column, such as 'Jokes of the Day' in *Lloyds' Weekly News,* featured approximately twenty jests in each weekly instalment — this quickly adds-up to 1,000+ jokes in a year, or 10,000+ over the course of a decade. Dozens, possibly even hundreds, of Victorian newspapers published columns of this nature every week, sometimes for several decades. Bestselling penny magazines, such as *Tit-Bits* and *Answers*, made even more extensive use of jokes; they often devoted their

front page entirely to jests and routinely sprinkled more of them inside. It was not uncommon for papers like this to feature between fifty and a hundred jokes in each weekly issue. Only a small portion of the press has been digitised. Nevertheless, if we process two decades of material from twenty of the most prolific digitised newspapers, we might expect to harvest something in the region of half-a-million jokes. Of course, many of them will be reprints — but, as previously explained, we consider the repetition of a joke to be a significant act, and each reprint gives valuable contextual data for interpreting its meaning and significance.

This does not mean that *TOJA* will be swamped by half a million reprints from *Punch*. Newspaper editors populated their joke columns from a variety of sources, including newspapers and magazines published overseas. As Nicholson has explored elsewhere, imported American jokes were hugely popular in Victorian Britain.[13] Editors gathered them into dedicated columns of 'Yankee Humour' which appeared everywhere from the biggest national papers to the smallest provincial weeklies. *Tit-Bits* magazine included plenty of transatlantic jokes, but also looked across the channel for fresh material; it's weekly page of 'Continental Tit-Bits' featured jokes clipped, and translated, from papers in mainland Europe. Other papers, such as the *Dundee Weekly News,* sourced material directly from their readers via long-running weekly joke competitions. The results for these competitions often printed the name and address of winning participants. When cross-referenced with the census, it is often possible to discover the age, occupation, and marital status of the people who submitted them — a rare, and extremely valuable insight into the public's joke-telling preferences.[14] In short, even though Victorian newspapers might not seem particularly humorous at first glance, they hold the greatest volume and variety of preserved jokes for this period *and* provide us

with the richest contextual metadata.

Unfortunately, extracting jokes from newspapers presents the biggest curatorial and technological challenge. After all, the vast majority of the articles in these publications are not suitable for our archive — jokes appeared alongside news stories, editorials, adverts, letters, stock prices, gardening columns, serialised stories, sports results, and dozens of other journalistic genres. At present, this is one of the chief obstacles faced by researchers who might want to explore the history of joke-telling in nineteenth-century Britain. It is not difficult to find Victorian jokes, once you know where to look, but it is much more challenging to find jokes about a specific pre-determined topic. Searching a digital archive of Victorian newspapers for the term ‘Gladstone’ will return plenty of jests about him, but these results will be overwhelmed by thousands of hits from other journalistic genres. Similarly, if we wanted to explore Victorian jokes made at the expense of lawyers or doctors, we would need to wade through millions of other articles that cite these professions. For instance, a search for the word ‘doctor’ in one nineteenth-century newspaper archive returns more than five million hits; if ten thousand doctor jokes appeared in these results, they would represent no more than 0.2% of the total. Moreover, our search would only find doctor jokes in which the word ‘doctor’ appears; a jest that signals its topic by naming a character ‘Dr Sawbones’ would remain undiscovered. This approach is also vulnerable to errors in an archive’s Optical Character Recognition (OCR) data, where the word ‘doctor’ is often mistranscribed by the software as ‘dactor’ or ‘docton.’ It is possible to construct complex queries that increase the percentage of useful hits, but the success of this approach is heavily dependent on an archive’s search tools and requires a detailed understanding of how joke columns were formatted. We developed *TOJA* to solve this problem and make historical

humour more accessible to non-expert researchers, but to accomplish this we needed to develop a strategy for extracting hundreds of thousands of jokes from millions of pages of digitised newspapers.

The easiest jokes to extract are those which were organised into consistently titled, long-running, weekly humour columns. Our dataset of nineteenth-century newspapers includes a 'document title' field in its metadata, which allows us to isolate individual columns based on the words that appear in their headers. For instance, we can extract all columns titled 'Jokes of the Day' that appeared in *Lloyd's Weekly News* – a total of 340 columns, each containing between 10 and 30 jokes. The same approach nets us 180 instalments of 'Jonathan's Jokes' from the *Hampshire Telegraph*, 798 columns of 'Wit and Humour' from the *Lancaster Gazette*, 102 editions of 'Clippings from the Comic Papers' published by the *Hull Packet*, and many more. Once the title of a long-running column has been discovered, it is relatively straightforward to extract all of its instalments from our newspaper dataset; at this stage of the curatorial process, a batch of 500 columns is no more difficult to gather than a set of five. Things become trickier for newspapers with a less settled format. The *Hampshire Telegraph,* for instance, changed the name of its imported American humour column on an almost weekly basis and resorted to the use of increasingly obscure transatlantic reference points; a column titled 'Buckwheat Cakes' sounds more like a recipe than a collection of jokes and is unlikely to be discovered via keyword searching. In cases like this, we have to gather columns manually at a slower pace. We also resort to manual methods in order to plug gaps caused by OCR errors. If a column was mistranscribed as being titled 'Jakes of the Doy' then its absence is often easy to identify simply by looking for unexplained interruptions in a column's weekly pattern.

Focusing on the titles of long-running columns is the most efficient way to extract jokes from nineteenth-century newspapers, but we know that this will not find *all* of the jokes hidden in our dataset. Many of them were printed outside of dedicated humour columns, sometimes in broader collections, and often as isolated column-fillers with no identifying title or recurring pattern. While it would technically be possible to browse through each newspaper, page by page, in search of every joke, this would be a laborious process for our small team to undertake. We are currently experimenting with two potential solutions. The first involves a crowdsourcing platform discussed in more detail in part two of this chapter. In brief, users of the archive and a team of volunteer 'joke detectives' can find and submit potential new sources to *TOJA* — subject to copyright, and the approval of the project team — in return for being credited as the joke's discoverer. The results of this process are unpredictable, and dependent on the unpaid labour of volunteers, but they have already identified viable new collections. Our second experiment uses a combination of Computer Vision (CV) and Natural Language Processing (NLP) to automatically identify potential jokes. The underlying algorithms are relatively complex but, in layman's terms, we feed a 'training corpus' of verified jokes into the system and instruct it to find articles that it thinks share similar visual and linguistic properties. Humour columns of the type discussed earlier tend to have characteristic visual patterns, such as shorter paragraphs and a repetitive layout. This visual search provides an initial filter, which results in a significant number of incorrectly identified blocks of text. In the second step, word and punctuation patterns characteristic to humour columns are used to filter out most of the incorrectly identified text blocks. For instance, if the characters '—*Punch*' or '—*Fun*' appear nearby, then a body of text is likely to be part of a joke column (the italics and dash

were typically used to signal an attribution). The resulting list of potential humour columns is then manually verified before being ingested into the system. The quality of the process and the amount of manual intervention needed depends heavily on the amount of training data available. As the archive grows, the amount of training data also grows, thereby gradually training the system to recognise jokes more effectively. Nevertheless, the manual verification step will always be required, as the algorithms are tuned to fish with a wide and fine net, as we believe it is preferable to discover as many jokes as possible, at the cost of finding and filtering out a lot of incorrect data. Both of these approaches are still under development but, when used in combination, we are optimistic that crowdsourcing and automatic identification will allow us to find and extract hundreds of thousands of long-forgotten jokes that would otherwise remain hidden.

Manuscripts and Other Fragments

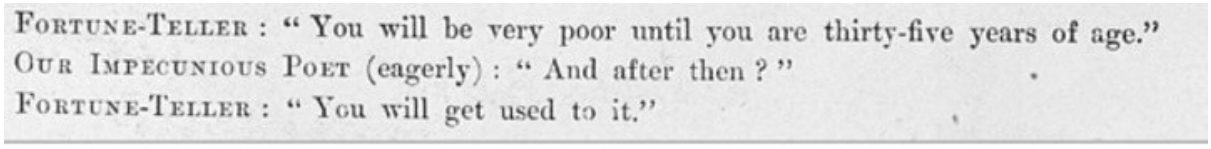
The majority of jokes in *TOJA* come from published sources — chiefly jest books, comic periodicals, and newspapers. However, traces of historical humour have also survived in an eclectic range of other documents and objects. Some people recorded jokes in their diaries, others jotted them down on postcards, and some scrawled them on toilet walls. The most exciting one-off sources we have discovered so far include the extensive and carefully organised files of Cliff Dean, a comedian from Lancashire operating in the 1940s, and a Victorian notebook filled with 91 pages of handwritten jokes assembled from a range of printed and oral sources. Our archive is receptive to this kind of evidence, but it is almost impossible to develop a systematic strategy for finding and processing it. We perform regular searches of

auction sites, second-hand bookshops, manuscript dealers, and archive catalogues. However, our best sources have so far been brought to us by people who learned about the project and wanted to contribute. We suspect there must be many joke-filled scrapbooks, notepads, postcards, and other ephemeral fragments sitting in people's attics all over the world. Over time, we hope that *TOJA* will continue to act as a magnet for this material and allow us to make it accessible to researchers for the first time.

Cataloguing Jokes

Once jokes have been located and extracted from their original documents, they must be processed and catalogued. This process consists of the following steps: splitting apart the individual jokes in an image; transcribing the image into text; and annotating the text. The workflow always uses an automate-first, manually-check/correct-after approach. An initial version of the output for each step is generated automatically. The result is then presented to volunteers through our purpose-built crowdsourcing tools. Depending on the quality of the automatic result, the volunteer may only have to verify its correctness. However, if the output contains errors, they may have to correct these manually or, in the worst-case scenario, repeat the whole step themselves. A detailed account of each step in this process lies beyond the scope of this chapter. However, in brief, images featuring multiple jokes — such as newspaper columns — are first broken down into clippings of individual jokes. This work is performed using a web-based tool which prompts users to draw a box around every joke; a similar approach is used by many citizen science projects, including those hosted on the industry-standard Zooniverse platform.[15]

Next, each of these clippings is automatically processed using OCR software, and these digital transcriptions are manually corrected by project volunteers whose contributions are automatically cross-checked for discrepancies and errors. Finally, volunteers and members of the core project team annotate each joke with metadata describing aspects of its format and focus. This metadata falls into three main groups: (1) the subject(s) of a joke; (2) the genre, or format, of a joke; and (3) subject/format specific details, such as information about a joke's characters and locations. Fig 1 outlines the metadata captured for a typical nineteenth-century 'dialogue' joke:



FORTUNE-TELLER : "You will be very poor until you are thirty-five years of age."
OUR IMPECUNIOUS POET (eagerly) : "And after then ?"
FORTUNE-TELLER : "You will get used to it."

Source Title: *The Sketch*
Source Type: Newspaper
Pub Date: 10 Oct 1894
Pub Location: London
Archive: BNA

Format: dialogue
Subjects: Poetry; Poets; Fortune Telling; Poverty.
Location: Unknown

Character 1: 'Fortune-Teller'
Gender: Unknown
Ethnicity: Unknown
Life-stage: Adult
Profession: Fortune-Teller

Character 2: 'Impecunious Poet'
Gender: Unknown
Ethnicity: Unknown
Life-stage: Adult
Profession: Poet

Fig 1: a nineteenth-century dialogue joke with metadata

Our project is not the first to wrestle with the challenge of categorising and organising jokes. Once again, we looked to the past for inspiration. The editors of the jestbooks in our collection all had to decide how their anthologies would be structured, and they reached a range of contrasting conclusions. Many, including the aforementioned *How To Set The Table In A Roar!,* eschewed any kind of organisation and printed their jokes in a seemingly random order. These unstructured books are well-suited to casual browsing, but do not function very successfully as reference works; if we want to find a joke about a particular subject,

we have to skim through the book until we find one. At the other end of the spectrum, some jokebooks organised their material thematically by subject. For instance, *Bennett Cerf's Vest Pocket Book of Jokes* (1957) is arranged into 135 alphabetised categories; under 'd' in the contents page we find dentists, department stores, dinner parties, doctors, dogs, and drink. The specificity of these categories varies greatly from one collection to another. Geoff Tibballs' *Mammoth Book of One-Liners* (2012) features separate categories for cats, dogs, ducks, and even turtles; in the personal joke files of the Lancashire comedian, Cliff Dean, this material was grouped under the broader category of 'Animals'.[16] At a more conceptual level, critics and comedians throughout history have asserted, with varying degrees of seriousness, that every joke can be placed within a small handful of broader, universal categories — unfortunately, each version of this theory confidently asserts a different set and number of categories. In short, there is no universally agreed system for classifying jokes, and we had to develop a robust taxonomy that would best fit the data and curatorial processes involved in the archive.

We attempted to establish a subject classification system at the start of our project, but it became apparent that no amount of forward-planning could anticipate all the jokes we later discovered. Wister's complaint that her contemporaries were happy to 'make a joke of anything' rings true. Some categories — marriage, lawyers, doctors, mothers-in-law, etc. — loom consistently large, but others ebb and flow in response to historical events and societal changes. A taxonomy developed and tested using Victorian humour would struggle to represent the comic culture of other periods and locations. However, one of the true benefits of digital archives is that it is easily possible to expand any categorisation and to categorise objects in multiple ways at the same time. To that end, our classification model is capable of expanding

in response to the discovery of new jokes. If a volunteer cannot find an appropriate classification in our system when they are annotating a joke, they can propose a new one, subject to the approval of the archive's curators. In many cases, these emergent topics can be flagged as sub-categories of existing subjects; jokes about motorcars, for instance, are a distinctive new presence at the turn of the twentieth century and deserve their own subject classification, but this sits within a broader topic of 'Transportation' which includes earlier gags about railways and cyclists. Crucially, each joke in *TOJA* can be annotated with multiple topic keywords, which solves a problem faced by the editors of printed jestbooks who usually needed to place a joke in a single category. Should a gag about airline food appear in the food section, or in the air travel section? In a digital archive like *TOJA*, it can be assigned to both. This also makes it possible for jokes within *TOJA* to be re-classified with new layers of descriptive metadata, perhaps in response to the development of alternative joke taxonomies, or in order to answer the research questions of other projects.

The jokes in *TOJA* are also categorised by genre/format. This was a common organisational strategy used by editors of joke books. Some, such as *Four Hundred* Laughs (1901), did not sort by subject, but devoted separate chapters to conundrums, verses, and witty sayings. Similarly, Cliff Dean's joke folders feature separate sections for comic songs and humorous newspaper advertisements. Devising an exhaustive taxonomy of joke genres presents the same methodological problems as categorising subjects; new genres and sub-genres emerge as our search broadens to new periods and locations, and many jokes do not fit into a single distinctive format. Once again, we have adopted a flexible approach that will allow us to expand, and refine, our categories as *TOJA* develops. We begin by

flagging the presence of a particular comic format or element — what might be termed its stylistic building blocks. For instance, most of the jokes in our Victorian pilot corpus currently feature one, or more, of the following elements: puns; dialogue; prose; verse; and Q&As. Next, we identify whether a joke belongs to a specific identifiable genre, such as a ‘conundrum’ or a ‘knock-knock’ joke. At present, this classification work is performed manually by volunteers using the project’s crowd-sourcing tools. However, we have begun to experiment successfully with computer-assisted classification. Genres with a very specific format, or distinctive linguistic markers (such as the quotation marks in a dialogue joke), can be identified automatically. For now, this works as a prompt to human users of the archive who are asked to verify the accuracy of the automated classification.

Finally, the metadata attached to each item in *TOJA* is shaped by the genre and content of the joke in question. For instance, the joke in Fig 1 was categorised as featuring ‘dialogue’, and so its cataloguer was prompted to provide additional information about the joke’s characters. Annotations like this are particularly important for improving the discoverability of jokes. Vital elements of a joke’s subject are often implied by context, or communicated playfully, instead of being signalled by a straightforward keyword. An undertaker might be referred to simply as ‘Mr Freshgrave;’ a man might refer to his mother-in-law as ‘your grandmother’ in conversation with his children; and an Irishman’s nationality might be signalled through dialect. None of these would be returned by a simple keyword search for ‘undertaker’, ‘mother-in-law’, or ‘Irish’. The richer we make this character metadata, the more nuanced our search queries can become. For instance, by capturing (wherever possible) the gender, profession/role, life-stage, and nationality of characters in a dialogue joke, we can focus a keyword search on passages of dialect

spoken by adult women, or identify gags in which a man converses with a child. However, there are limits to the depth of detail we can achieve. At present, we can only capture unambiguous aspects of a character's identity. We initially hoped to record the social class of a character, but found that this could not be inferred reliably by non-expert volunteers; fortunately, the 'profession/role' category acts as an alternative way to explore this issue. Similarly, if a joke published in London features a man speaking Standard English, then we might reasonably assume that he was Englishman — but we do not record this character's nationality unless it is signalled more directly. In all cases, if no evidence exists to support an annotation — if a character's gender, for instance, is not stated or very strongly implied — then it is left blank. For instance, the implied genders of the characters in Fig. 1 are female (fortune-teller) and male (poet), but this is not signalled decisively enough to become part of the joke's factual metadata.

Built on top of the images, texts, and meta-data is an interface for searching and exploring the archive, which aims to support both focused search and open-ended exploration and browsing in the archive. The search interface uses a standard faceted-search interface, which enables the researcher to input keyword searches or construct more complex queries such as all verse jokes printed in books featuring a teacher, a pupil and the word 'apple'. It also provides access to the annotated texts in TEI format, which enables offline working with the joke corpus and integration with other Digital Humanities research tools.[17] While *TOJA* is partly intended to facilitate complex, academic research queries, we also recognise that jokes have a leisure life; that users might browse them for pleasure in a much less focused or analytical way. To support this, the archive provides an interface that allows the user to choose from a range of automatically and manually curated topics and then explore, read,

and enjoy the jokes.

Conclusion

At the time of writing, the first prototype of *The Old Joke Archive* is nearing completion and will soon be ready for public testing. Once launched, it will be possible for researchers to access tens of thousands of long-forgotten jokes and explore them with a newfound precision. It is too early to predict what impact this might have on the ways that we explore and understand the history of joke-telling. However, the construction and design of the archive has already helped us to understand the complex place of jokes in the historical record. Despite their inherent ephemerality, it turns out that nineteenth-century jokes *have* survived in remarkably large numbers — and the same is likely true of many other historical periods and places. Millions of jokes have been preserved in archives and personal collections around the world. They have been captured in the pages of jestbooks, comic periodicals, newspapers, and manuscripts. Sometimes this preservation was deliberate and systematic but, more often, it was haphazard and accidental. Many jokes have survived as stowaways in the margins of more valued texts; as column-fillers that were used to plug gaps between newspaper articles, or throwaway quips in diaries. The diffuse nature of this historical record has long been an obstacle to researchers and has tended to direct our attention to the most prominent and accessible sources of material. *TOJA* aims to redress this imbalance, but gathering and cataloguing jokes from such a wide array of unidentified sources presents significant curatorial challenges. As this chapter has outlined, the solution involves a combination of historical expertise, computer science, and the generous help of

volunteers. There is much more work to be done in each of these areas, and millions more old jokes to be found. Over the coming years, we hope that fellow researchers in the field of humour studies will join us to expand the temporal, linguistic, geographical, and curatorial limits of *The Old Joke Archive* and uncover new insights into the forgotten humours of the past.

BIBLIOGRAPHY

Anon. *How To Set The Table In A Roar! The jester's, punster's and humourist's guide.* London: Miller and Co., c.1880s.

Cerf, Bennet. *Bennett Cerf's Vest Pocket Book of Jokes*. London: Hammond, Hammond & Company, 1957.

Dickie, Simon. *Cruelty and laughter: Forgotten Comic Literature and the Unsentimental Eighteenth Century*. Chicago: Chicago University Press, 2011.

Gray, Donald J. 'A List of Comic Periodicals Published in Great Britain, 1800-1900, with a Prefatory Essay'. *Victorian Periodicals Newsletter* 5, no. 1 (1972): 2-39.

Kemble, John R. *Four Hundred Laughs: or Fun Without Vulgarity*. New York: A. L. Burt, 1901.

Nicholson, 'The Victorian Meme Machine: Remixing the Nineteenth-Century Archive'. *19: Interdisciplinary Studies in the Long Nineteenth Century* 21 (2015).

Nicholson, Bob. '"You Kick the Bucket; We Do the Rest!": Jokes and the Culture of Reprinting in the Transatlantic Press'. *Journal of Victorian Culture* 17, n. 3 (2012): 273-286.

Nicholson, Bob. 'Capital Company: Writing and Telling Jokes in Victorian Britain'. In *Victorian Comedy and Laughter: Conviviality, Jokes and Dissent*, edited by Louise Lee. London: Palgrave, 2020.

Nicholson, Bob. 'Jonathan's Jokes: American Humour in the late-Victorian Press'. *Media History* 18, no. 1 (2012): 33-49.

Reinke-Williams, Tim. 'Misogyny, Jest-Books and Male Youth Culture in Seventeenth-Century England'. *Gender and History* 21, no. 2 (2009): 324-339.

Tibballs, Geoff. *The Mammoth Book of One-Liners*. London: Constable & Robinson, 2012.

Verberckmoes, Johan. *Laughter, Jestbooks and Society in the Spanish Netherlands*. London: Palgrave, 1999.

Wister, Sarah Butler. 'A Plea for Seriousness'. *Atlantic Monthly* (May, 1892): 625-630.

[1] Wister, ‘A Plea for Seriousness’. This article was initially published anonymously, but it was subsequently attributed to Wister in the *Atlantic Monthly’s* annual index.

[2] Ibid., 628.

[3] Ibid., 628.

[4] Ibid., 625.

[5] Ibid., 627.

[6] See Nicholson, ‘Capital Company’.

[7] See, for example: Dickie, *Cruelty and Laughter*; Verberckmoes, *Laughter, Jestbooks and Society*; Reinke-Williams, ‘Misogyny, Jest-Books and Male Youth Culture’.

[8] Anon., *How To Set The Table in A Roar!*

[9] Nicholson, ‘You Kick the Bucket’.

[10] Gray, ‘A list’.

[11] Nicholson, ‘Capital Company’.

[12] Nicholson, ‘The Victorian Meme Machine’.

[13] Nicholson, ‘Jonathan’s Jokes’.

[14] Nicholson, ‘Capital Company’.

[15] The Zooniverse Project, www.zooniverse.org

[16] The joke files of Cliff Dean are privately held by the Dean family.

[17] Text Encoding Initiative, https://tei-c.org.