The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP with 180,000 Page Images

William Ughetta
Department of Computer Science
Princeton, New Jersey, USA
wughetta@cs.princeton.edu

Brian W. Kernighan
Department of Computer Science
Princeton, New Jersey, USA
bwk@cs.princeton.edu

ABSTRACT

The Proceedings of the Old Bailey is a corpus of over 180,000 page images of court records printed from April 1674 to April 1913 and presents a comprehensive challenge for Optical Character Recognition (OCR) services. The Old Bailey is an ideal benchmark for historical document OCR, representing more than two centuries of variations in documents, including spellings, formats, and printing and preservation qualities. In addition to its historical and sociological significance, the Old Bailey is filled with imperfections that reflect the reality of coping with large-scale historical data. Most importantly, the Old Bailey contains human transcriptions for each page, which can be used to help measure OCR accuracy. Since humans do make mistakes in transcriptions, the relative performance of OCR services will be more informative than their absolute performance. This paper compares three leading commercial OCR cloud services: Amazon Web Services's Textract (AWS); Microsoft Azure's Cognitive Services (Azure); and Google Cloud Platform's Vision (GCP). Benchmarking involved downloading over 180,000 images, executing the OCR, and measuring the error rate of the OCR text against the human transcriptions. Our results found that AWS had the lowest median error rate, Azure had the lowest median round trip time, and GCP had the best combination of a low error rate and a low duration.

CCS CONCEPTS

• **Applied computing** → **Optical character recognition**; • **General and reference** → **Performance**; **Evaluation**.

KEYWORDS

Optical Character Recognition, Old Bailey, Historical Documents, Amazon Web Services, Microsoft Azure, Google Cloud Platform

ACM Reference Format:
William Ughetta and Brian W. Kernighan. 2020. The Old Bailey and OCR: Benchmarking AWS, Azure, and GCP with 180,000 Page Images. In *ACM Symposium on Document Engineering 2020 (DocEng '20), September 29-October 2, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3395027.3419595

DocEng '20, September 29-October 2, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-8000-3/20/09…$15.00
https://doi.org/10.1145/3395027.3419595

1 INTRODUCTION

The Proceedings of the Old Bailey is a collection of final words at Tyburn and trial testimonies at the Old Bailey courthouse in London between April 29th, 1674 and April 1st, 1913 [2]. The Old Bailey documents have over 180,000 pages of testimonies, accounts, verdicts, and sentences, covering almost 240 years. Historically, the Old Bailey only published a small portion of the actual number of trials, which were hand-picked for entertainment and to promulgate a narrative of justice [6].

The Old Bailey Online project, directed by Clive Emsley, Tim Hitchcock, and Robert Shoemaker, digitized the Old Bailey documents over the course of eight years, starting in 2000 [4]. The goal of this paper is to benchmark AWS, Azure, and GCP's cloud OCR services against the Old Bailey's human transcriptions.

AWS, Azure, and GCP have been compared before (on at least 3 images) [3], and the Old Bailey has been used before to measure OCR (with 20 images) [1]; we believe this paper is the first to benchmark AWS, Azure, and GCP on the entire Old Bailey corpus.

The Old Bailey is composed of two smaller publications. The first, Ordinary's Accounts (OA), carried the last words of capitally-convicted criminals and was published from May 17th, 1676 to October 14th, 1772. There are 4,321 images in OA (excluding 33 missing images). The second publication is Sessions Papers (SP), which carried the court proceedings and was published from April 29th, 1674 to April 1st, 1913. There are 179,056 images in SP (excluding 361 missing images). To evaluate accuracy, we used the human transcriptions of each page image to compute the Character Error Rate (CER) of each OCR result. CER is the Levenshtein distance [5] between the OCR and the human transcription divided by the number of characters in the human transcription.

The Old Bailey Online claims an "accuracy well over 99%" [2] and, even if the accuracy was well below 99%, the benchmark would still be fair to OCR providers on a relative basis. The Old Bailey certainly has character. Examples include the hole shown in Figure 1, a quote in Latin in Figure 2, and many more variations from blurry, white, or black scans to photographs with fingers on the edge. The Old Bailey has an impressive range of printing and preservation qualities. Additionally, humans transcribed the meaning of the text, not what is written on the page. Examples of this include omitting page numbers and hyphenated suffixes of words in the bottom right corner of a page. In one instance, the transcriber introduced an error by transcribing a well-intentioned annotation in the acquittals list, resulting in two names being combined as one "Dean Dealler."

The Old Bailey Online project transcribed the Old Bailey in two stages. The first stage, from 2000 to 2005, digitized all 4,354 Ordinary's Accounts and 58,699 Sessions Papers, which are all the Sessions Papers up to and including October 1834. In this first group,

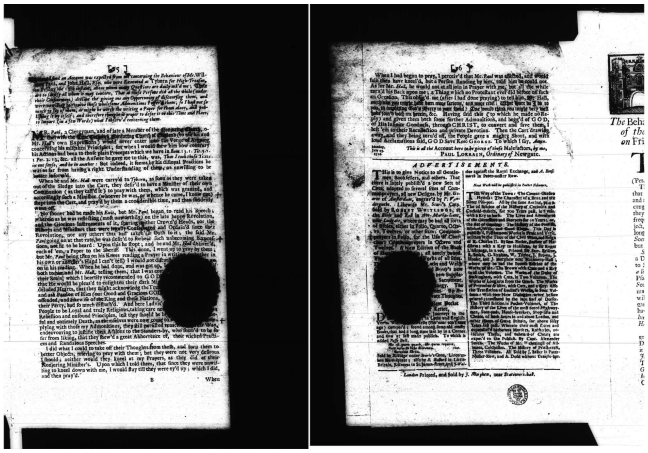

Figure 1: Ordinary's Accounts July 13th, 1716 (# 625-6)
Character Error Rate **Left: AWS 12.31%; Azure 61.24%; GCP 10.95%**
Right: AWS 58.16%; Azure 80.46%; GCP 59.19%

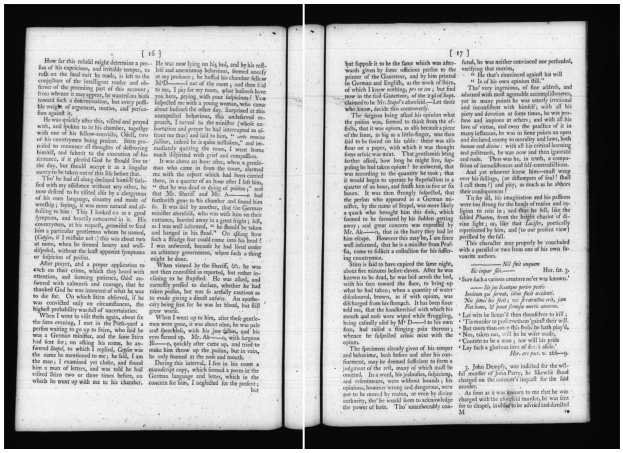

Figure 2: Ordinary's Accounts September 15, 1760 (# 3421-2)
Character Error Rate **Left: AWS 73.97%; Azure 35.26%; GCP 76.64%**
Right: AWS 75.90%; Azure 12.17%; GCP 77.36%

each page was transcribed independently by two humans, and then the results were merged together.

The second stage, from 2006 to 2008, totaled 120,660 images, and represented the remainder of the Sessions Papers. Images in this second group were transcribed by one human and one OCR pass before being merged [2]. The image counts were calculated by parsing the Old Bailey Online's XML files, which contained image pointers as well as human transcriptions with metadata tags. The 58 image pointers that did not have human transcriptions were not included in the count. All images are in JPG format except Sessions Papers before January 1834, which are in GIF format.

2 METHODOLOGY

Benchmarking AWS, Azure, and GCP involved three main steps: acquiring the Old Bailey dataset, executing over 180,000 OCR jobs on each provider, and measuring the accuracy of the results. The code and data is available at https://github.com/ughe/old-bailey.

2.1 Data Acquisition

We first parsed the Old Bailey Online XML files to obtain image pointers, which were then used to download individual images. While checking the download, we found and were able to fix 140 broken image pointers. We have submitted corrections to the Old Bailey Online project so that others will be able to find the 140 images once the XML has been updated. Another 430 pointers failed to return anything; we believe those images are simply missing.

For as large as the Old Bailey is, the images average only a few hundred kilobytes each, totaling about 32GB altogether. This small size made all of the subsequent operations significantly faster. Images range from black and white to color; from single page scans to double; and from as little as 63KB to as large as 1.9MB (or 2.7 MB when all GIFs are converted to JPGs). The smallest image (pointer 18) has a DPI of 300, and is large enough for a human to read clearly.

In total, there were 183,771 pointers parsed from the original XML. Only 183,713 pointers had human transcriptions and of those only 183,434 images downloaded successfully.

2.2 OCR Execution

We created a tool to simplify running OCR on AWS, Azure, and GCP and to output the resulting text and the round trip time in a common JSON format. The tool also saves the original raw response so that information such as coordinates or confidence levels could be used in future research without re-running all 180,000 images.

We aimed to use the default parameters for each OCR service as much as possible. Specifically, we set AWS to re-try three times before failing; Azure to detect language but not page orientation; and GCP did not have any parameters. We only used the synchronous APIs, which simplified the client code, enabled more accurate timing measurement, and saved money by uploading images at run-time instead of storing them for asynchronous processing in S3-style buckets. Both AWS and Azure returned lists of regions, lines, and words, which needed to be manually chunked together. We did so without any attempts to detect columns or to use the coordinates and confidence levels. GCP was the only provider that also returned a full text transcription automatically.

Running the OCR was done between May 5th and May 10th, 2020. First, we processed all 4,321 Ordinary's Accounts images. The Ordinary's Accounts were processed on AWS, Azure, and GCP in less than 6.5 hours between May 5th and 6th. Next, we ran the 121,055 Sessions Papers that were in JPG format using four parallel bash scripts. The remaining 58,001 images, all Sessions Papers before 1834, were converted from GIF to JPG format (AWS did not support GIFs), taking about 3 hours, and then run through OCR. As soon as we started the GIF process, AWS began to rate limit us, while Azure and GCP sailed on without issue. We immediately halted one of the four JPG groups. On May 8th, about 38 hours from the start, the first JPG group finished. An hour later the next two JPG groups finished. At this point, we restarted the one halted JPG group since AWS could now cope with the reduced rate. This last JPG process then finished about 30.5 hours later, which was on May 9th. The only process still running was the GIF group of 58,001 images (converted to JPG). These images finally finished on May 10th, which was approximately 97.5 hours from their start.

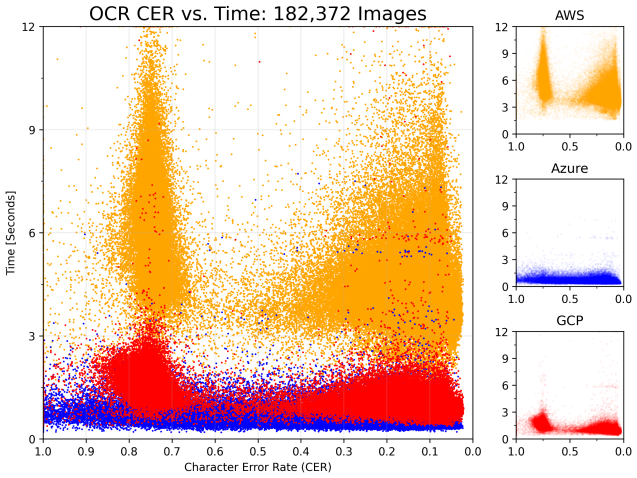

Figure 3: OCR Character Error Rate vs. Time in Seconds

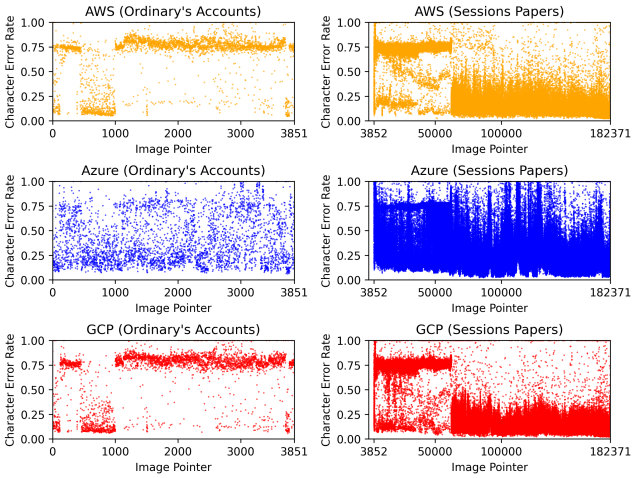

Figure 4: OCR Character Error Rate: 182,372 Images

2.3 Accuracy Measurement

The Character Error Rate (CER) is a proxy for OCR accuracy. As detailed in Section 1, the CER is the Levenshtein distance between the OCR result and the human transcription divided by the number of characters in the human transcription. The Levenshtein distance is the minimum number of edits—insertions, deletions, or substitutions—required to transform one string of text into another. There are pages in the Old Bailey where transcribers do not have any text for them, yet the physical pages do have text. In the event that the human transcription has no text, then the CER is defined to be 100% if the Levenshtein distance is non-zero and 0% otherwise.

Once the Levenshtein distances and the CER were calculated, we compiled all of the results into a single CSV file for creating graphs and analyzing the results.

Finally, we only compared results if all three services ran the OCR successfully. Other reasonable approaches would have been to count partial results or to re-run the test again on all three providers if the error was a timeout.

3 EVALUATION

Although we ran OCR on 4,321 OA images and 179,056 SP images, we only used results that were returned correctly from all three providers. Only 3,852 OA images and 178,520 SP images were returned correctly from all three services.

We found that AWS had the lowest median error rate at 17.3%, Azure had the lowest median round trip time at 0.492 seconds, and GCP had the best trade-off between speed and error rate, which were 0.998 seconds and 19.7% CER.

3.1 Character Error Rate vs. Round Trip Time

Figure 3 puts the performance in context of speed by plotting the CER versus the round trip time for each of the three cloud providers. The graph thresholds both the maximum time at 12 seconds and the maximum error rate at 100%. The optimal accuracy and speed is in the lower right-hand corner of each square graph. The worst case of slow processing and a wrong result is in the upper left-hand corner of each. Each plot for AWS, Azure, and GCP is also shown independently to the right in Figure 3 since GCP is drawn on top of Azure, which is drawn on top of AWS.

Figure 3 shows that AWS spends the most time with either a low error rate or an extremely high error rate–without many results in between. Azure holds a disciplined line showing that there is not much of a correlation between accuracy and speed. GCP also has a bimodal appearance, but it is much more disciplined than AWS.

3.2 Character Error Rate

Figure 4 plots the character error rates for each of the 182,372 images across the three cloud providers. The first column in the figure shows 3,852 images from Ordinary's Accounts ranging from years 1676 to 1772. The second column shows 178,520 images from Sessions Papers and ranges from years 1674 to 1913. Note that all 3,852 images in Ordinary's Accounts would fit into the small white sliver to the left of the 3,852 tick in the Sessions Papers graphs. Additionally, AWS and GCP appear to have almost identical error rates on both the small scale of Ordinary's Accounts and the large scale of Sessions Papers.

In Figure 4, Ordinary's Accounts has a striking period of improvement in accuracy from about pointers 400 to 1000, which correspond with the dates September 1708 and May 1725, respectively. Azure does not display the same improvement over the same period. Figure 1 (pointer numbers 625 and 626) and Figure 2 (pointer numbers 3,421 and 3,422) reflect the graph. Figure 1, in the low CER range between 400 and 1000, does have a lower CER for AWS and GCP than the latter images that are outside the area of improvement. The character error rates are shown in the respective figures.

The most striking feature of the graphs in Figure 4 is the vertical line shown at approximately 62,500 for both AWS and GCP Sessions Papers. The pointer 62,500 corresponds to December 5th, 1834. This is only 255 pointers away from the final page of October 1834 (at pointer 62,245), which was the boundary between the first transcription group created from 2000 to 2005 and the second transcription group created from 2006 to 2008. The stark separation of

CER					
Provider	Min	Q1	Q2	Q3	Max
AWS	2.3%	9.5%	17.3%	73.2%	1226.7%
Azure	2.3%	15.2%	23.6%	38.8%	322.3%
GCP	2.1%	11.6%	19.7%	74.1%	1063.0%
Seconds					
Provider	Min	Q1	Q2	Q3	Max
AWS	1.425	3.871	4.505	5.634	42.066
Azure	0.196	0.428	0.492	0.620	7.719
GCP	0.376	0.856	0.998	1.344	116.207

Table 1: Character Error Rate & Time (Seconds) Quartiles

	AWS	Azure	GCP
OA Success	4,320	3,853	4,321
OA Errors	1	468	0
SP Success	178,567	178,871	178,834
SP Errors	489	185	222
Total OCR Success	182,887	182,724	183,155
Total OCR Errors	490	653	222
Cost per Page	0.0015	0.0015	0.0015
Expected Amount	274.33	274.09	274.73
Billed Amount	274.48	274.19	273.46
Billing Error	0.054%	0.038%	-0.463%

Table 2: Cloud Providers' Costs: Expected vs. Billed

AWS and GCP error rates mostly above 70% before pointer 62,500 and mostly below 30% after pointer 62,500 suggests that the CER is highly correlated with the divisions in preparing the dataset. Although Azure does not have as strong a change, a similar horizontal line around 75% is visible from pointer 3,852 to about 62,500.

If the correlation was caused by the preparation of the dataset, then one possible explanation is that humans may have made more mistakes in the first group than the human and the OCR made in the second group. Another explanation is that the documents in the first group are harder to decipher than the images in the second group. While Figure 1 shows two difficult images from the first group, Figure 2 shows an easier pair of images from the first group.

Surprisingly, AWS and GCP perform better on the harder images, while Azure performs better on the easier images. AWS and GCP achieve error rates as low as 12.31% and 10.95% in Figure 1 (without columns) and as high as 75.90% and 77.36% in Figure 2 (with columns). Conversely, Azure does better in Figure 2 with 12.17% CER and worse in Figure 1 with 61.24% CER for the same images.

3.3 Quartiles

Table 1 presents another approach to considering the OCR accuracy and speed with quartiles for each provider. AWS leads the median CER ranking followed by GCP and Azure. Although Azure has the worst median CER, it has the best third quartile CER performance of 38.8%. This compares to AWS's 73.2% CER and GCP's 74.1%.

Table 1 also shows the OCR round trip time quartiles for each provider in seconds. Azure had the best median duration with 0.492 seconds; GCP with 0.998 seconds; and finally AWS with 4.505 seconds. Notably, Azure and GCP's third quartiles (0.620 and 1.344 seconds, respectively) are both faster than AWS's fastest OCR duration of 1.425 seconds.

3.4 Cost

AWS, GCP, and Azure all charged $1.50 per 1,000 images of any size below their limits. Results are shown in Table 2. The expected amount for each provider is the product of the cost per page and the number of total successful calls. The costs show that cloud providers accurately charge for their OCR services, within half of a percent of the expected value. GCP was the only provider that charged less than the expected price, and Azure had the lowest billing error rate.

4 CONCLUSIONS

The Old Bailey presents a unique opportunity for evaluating OCR. From its variety of contents—including spellings, layouts, preservation, and scanning quality—to the essential human transcriptions, the Old Bailey offers a significant challenge for cloud OCR services. Some pages cannot be read at all. Others cannot be deciphered by most humans. Still others are clear as day. The Old Bailey's range of appearances from poor preservation, digitization, mistakes in the original contents of the documents, and even errors in the human transcriptions are a challenge for any OCR system.

AWS and GCP had the lowest median error rates of 17.33% and 19.74% respectively. Azure and GCP had the lowest median round trip times of approximately 0.5 and 1 second, respectively. These results are shown in Table 1. Perhaps the largest surprise was how similar some of the results were on a large scale. This could be an indicator either of the high degree of similarity between services or of the disparity between the contents of the page and the contents of the human transcription. If AWS, Azure, and GCP all recognized the same text that was different from the human transcription, then Azure, AWS, and GCP would be expected to have a similar error rate; however, Azure's CER does not track AWS and GCP results.

ACKNOWLEDGMENTS

The authors are grateful for help and advice from David Brailsford, Zoe LeBlanc, Sharon Howard, Tim Hitchcock, Mikki Hornstein, Chris Miller, Rasool Tyler, and Jack Brassil, and for funding from Princeton SEAS and GCP credits.

REFERENCES

[1] Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised Transcription of Historical Documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 207–217.

[2] Clive Emsley, Tim Hitchcock, and Robert Shoemaker. 2020. *Old Bailey Online*. Retrieved July 3, 2020 from www.oldbaileyonline.org Version 7.0.

[3] Bill Harding. 2019. *2019 Examples to Compare OCR Services*. Retrieved July 3, 2020 from https://www.amplenote.com/blog/2019_examples_amazon_textract_rekognition_microsoft_cognitive_services_google_vision

[4] Sharon Howard. 2017. Old Bailey Online XML Data. https://doi.org/10.15131/shef.data.4775434.v2

[5] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10 (1966), 707–710.

[6] Robert B. Shoemaker. 2008. The Old Bailey Proceedings and the Representation of Crime and Criminal Justice in Eighteenth-Century London. *Journal of British Studies* 47, 3 (2008), 559–580. http://www.jstor.org/stable/25482829