OCR and the Bleeding Obvious

Nick Reddan

Today an e-mail appeared in my inbox from someone who had found a newspaper extract entry on my website, (https://nickreddan.net). The extract was from *Pue's Occurrences* (Dublin)[1] in 1754. The reference was to 1 February 1754 and text is as follows:

> *Friday last died suddenly James STAUNTON, the elder of Youghall in the county Galway, Esq, a gentleman greatly regretted by all his acquaintances.*

The sender of the email was seeking confirmation of the information, or at least of the newspaper report.

These extracts were mostly gathered by me in the 1990s and 2000s from microfilms held in the National Library of Australia (NLA). The microfilm readers I used has simple optical interfaces that magnified the image on the microfilm to readable size. Since then optical character recognition (OCR) has been used to convert microfilms of newspapers to searchable databases that have proved a great boon for family history researchers.

The largest coverage of British and Irish newspapers is in the British Newspaper Archive based on the British Library microfilms. It has some coverage of Pue's. The Irish Newspaper Archive does not appear to have *Pue's Occurrences*.

I did a search on British Newspaper Archives but had no result for James STAUNTON in the 1750s. Thus, I thought it best if I went back to the source I had used originally, namely the microfilm in the NLA. Things have changed.

Now the image you see is a digital one that is sent to the computer screen from the scanner. This process is a little slower for browsing a range of dates but has the advantage you get an image file which you can save to a thumb-drive and play with at home.

After a bit of fiddling I went to the relevant issue of *Pue's Occurrences*.

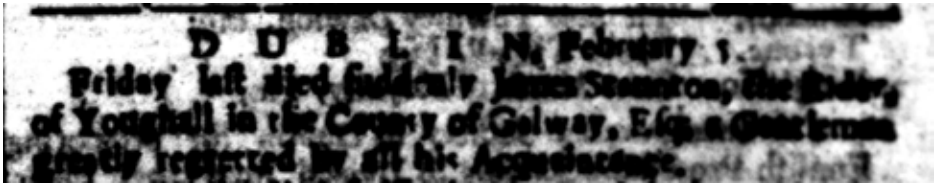

VOL. LI. NUMB. II.

PUE'S OCCURRENCES.

From SATURDAY February the 2d, to TUESDAY February the 5th, 1754.

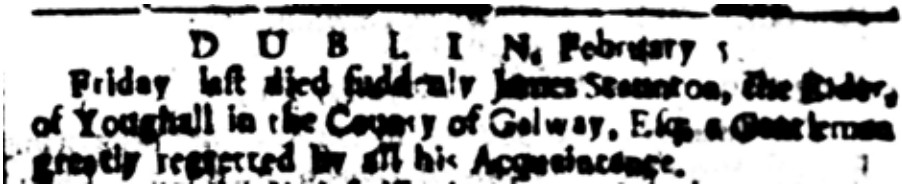
What you see is the original printed page has some bleeding of ink from the other side and the generous use of ink means some things are not very clear.

This is an unaltered TIF image produced by the scanner of the relevant entry which I had extracted.



February 1 is pretty clear but the rest needs a bit of interpretation to get the word out.

Even after some manipulation the image is far from clear:



The words in this version are reasonably clear. Note the long ‘s’ in “last” and “Suddenly”. Although “Staunton” is fairly rough knowing that STAUNTON is a Galway family it is pretty clear. “The Elder” is again not very clear but “younger” would be longer and have a couple of tails below the line.

Looking at these images you can understand why the OCR may have failed pick up “James Staunton, the Elder”.

Do not get me wrong, I love OCR newspaper sites and use them a lot. Where would we be without Trove?

This article is a small example to bring out the fact that OCR does not pick up everything. Sometimes, the hard work of poring over newspapers or microfilms provides some additional gems. I am thankful for the wonderful resources at the NLA and for the work of people like Rosemary ffolliott and H F Morris who collected and published great numbers of newspaper extracts.

1 From NLA Catalogue: *Pue’s occurrences [microform], mfm X 596*, Series: Irish newspapers in Dublin libraries, 1685-1754.