Filter bubbles

written by Jeremy Huggett | 20/02/2015

In an earlier post I wrote about the importance of understanding the legibility, agency and negotiability of archaeological data as we increasingly depend on online data delivery as the basis for the archaeologies we write and especially as those archaeologies show signs of being partly written by the delivery systems themselves.

A simple illustration of this is the idea of *filter bubbles*. This term was coined in 2011 by Eli Pariser to describe the way in which search algorithms selectively return results depending on their knowledge of the person who asked the question. It's an idea previously flagged by, amongst others, Jaron Lanier who wrote about 'agents of alienation' in 1995, but it came to the fore through the recognition of the personalisation of Google results and Facebook feeds (and is the counter-selling point of the alternative search engine, DuckDuckGo, for example). So can we see this happening with archaeological data? Perhaps not to the extent described by Pariser, Lanier and others, but still …

For example, here in Scotland we benefit from access to Canmore, the National Monuments Record and allied resources maintained by RCAHMS. Searching Canmore for '*motte and bailey*' castles and restricting sites to Scotland (to filter out some English sites returned via the Brian Hope Taylor collection held by RCAHMS) returns 41 hits – a mix of the great and the small, and including six 'possibles'.

Turning to the Archaeological Data Service's online UK-wide catalogue, we can use the ArchSearch browser to repeat the search. Selecting '*motte and bailey*' from the Monument Types > Defence > Castle category, and '*Scotland*' from the Where category, returns … 437 hits. Hmm, not what we might have expected!

Fortunately, the ArchSearch facetted browser gives a hint of what's going on – the results are including records from the West of Scotland Historic Environment Record which are duplicating records in Canmore. Adjusting the search to restrict it to only those records in Canmore (since the ADS holds a copy of Canmore) returns … 339 records. Closer, but still a long way to go. Visually scanning the records returned, one thing becomes clear – ArchSearch is using a fuzzy search. Although the search criterion was '*motte and bailey*', it's actually returning mottes as well as baileys and motte and baileys (mottes and baileys?). As a final throw of the die, adding '*bailey*' as a keyword to the ArchSearch criteria returns 48 hits, which seems hopeful.

At least that's down to the level where it's feasible to do a side-by-side comparison. This shows that of the 48 ArchSearch records, 8 are not returned by Canmore, while one of the Canmore hits isn't returned by ArchSearch. Of those 8, three are just baileys although one, Gallows Hill (ArchSearch; Canmore), is adjoining a motte so it might seem strange Canmore hasn't classified it as such (indeed, whether you can have a bailey without a motte is perhaps debatable). A further three of the 8 aren't in Scotland at all, and investigating further, two of these are records from the Brian Hope Taylor collection. Restricting the search criteria to Scotland removed these from the Canmore

results, but the same restriction applied to ArchSearch has not. The third of these sites, Liddel Strength (ArchSearch; Canmore), appears to be a victim of a border change between England and Scotland during the lifetime of the Scottish National Monuments Record and also appears in English Heritage's Pastscape. Coincidentally, the site that Canmore returns which isn't included in the ArchSearch results is another motte and bailey castle called Liddel in the same general area as Liddel Strength but a few miles into Scotland. Of the remaining two from the 8, both Hutton Mote (ArchSearch; Canmore) and Tinwald (ArchSearch; Canmore) are classified by Canmore as mottes rather than motte and baileys, although in each case a possible bailey is noted in the accompanying archaeological notes.

So in the end, we can finally match 40 of the records returned by both Canmore and ArchSearch in response to the same search, with one site missing entirely from the ArchSearch results. (For a variant on this exercise, see Huggett 2014).

So where does all this leave us? One thing is clear – there are discrepancies in both sets of results, and neither is complete, but then neither organisation would claim that their data are ready to be used as is, without some further processing. Indeed, these kinds of discrepancies in the data are precisely why there has been some reluctance in the past to make datasets such as these freely available online. More problematic is that two essentially identical searches generate widely differing results because the underlying search implementation is different despite surface appearances. There are filter bubbles of sorts in operation – for instance, the ArchSearch search is a fuzzy one despite the appearance of using a specific classification (underlined by the appearance of '*motte*' as a separate search criterion), and for some reason the geographical restriction to Scotland isn't comprehensively applied. Only a workaround got close to identical results. Unlike Pariser's filter bubble, two people performing the same search in Canmore or in ArchSearch aren't going to get different results, but one person searching Canmore and the other ArchSearch will, and indeed, a casual search of either will throw up incomplete and/or incorrect results unless effort is taken to understand what has been returned.

None of this is to criticise either data provider – pick any other large provider of similar data and the same kinds of issues are encountered. This is what happens when data collected over a long period of time, by different people, subject to different processing, and presented via different interfaces and search tools, are made available. And better by far that the data are available for us to work with! But what this does demonstrate is the challenges entailed in making accurate and reliable data available, and the risks inherent in combining and linking these kinds of data using automated techniques such as natural language processing. In some respects, it is better to approach these data as if they are not (yet) data at all – to recognise that in various ways they are contaminated by methodological, theoretical, human and technical bias, that they are partial and selective, and that they are constrained by both the conditions of their creation and their mode of delivery. It underlines that sometimes data cannot be considered to ***be*** data until they are cleaned and verified – and how much cleaning and verification is necessary before they can be considered usable data is a question of individual judgement and experience.

References

E. Pariser 2011 *The Filter Bubble: What The Internet is Hiding From You* (Penguin, London)

J. Huggett 2014 ‘Promise and Paradox: Accessing Open Data in Archaeology‘. In: C. Mills, M. Pidd and E. Ward (eds.) *Proceedings of the Digital Humanities Congress 2012*. Studies in the Digital Humanities. Sheffield: HRI Online Publications.