Jane Winters: Too much information? - Negotiating the archives of the Web

For historians, and researchers in many other humanities disciplines, web archives remain a largely unknown, and certainly underused, primary source. Even within digital humanities, web archives as a focus for analysis have remained on the fringes. It is, however, increasingly hard to imagine how you might study the history of the developed world in the late 20th and early 21st century without turning to the archived web. Web archives are the big data that it will be impossible for historians to ignore, but they pose a formidable set of challenges, ranging from the technical to the legal, and all points in between. The novelty of these challenges is sometimes overstated, but the scale and heterogeneity of web archives can seem overwhelming.

This presentation will discuss the difficulties of working with the archived web, using the .uk domain as a case study. There is no single archive of the UK’s historical web, rather there are many archives, which overlap and diverge in multiple and largely unknown ways. The British Library alone has three separate collections of web archives: data purchased from the Internet Archive for the period from 1996 to April 2013, which is fully searchable; material crawled from the web since April 2013 in accordance with legal deposit legislation, to which there is only limited on-site access; and the open, but selective UK Web Archive. Defining the relationships between the multiple archives of UK web space will be essential for our understanding of the possible shape(s) of a national web sphere, and is a necessary first step to more sophisticated quantitative and qualitative analysis. The presentation will conclude by considering the extraordinary richness of web archives for humanities research, and why we should take the time, and make the effort, to understand how they are constructed and what they contain.