Lou Burnard (TEI, Humanities Computing, Oxford/Paris)

Language as Data in the Humanities (Abstract)

Epistemological considerations about knowledge, justification, and the rationality of belief almost always have recourse to considerations of something called "data": we say "this is data driven" in a positive sense, or "show me the data" in a negative one. We talk approvingly of data-driven learning, and of evidence-based medicine. We systematically claim to prefer the empirical to the "armchair" variety of linguist.

But as Stanley Fish pointed out a very long time ago, no matter how objective our data-gathering methods, there comes a point at which the data has to be interpreted. And let us not also forget that the data we choose to analyse in the humanities is itself socially and historically determined: vast as they are, the digital collections of Hathitrust or Googlebooks or Gallica or represent in their composition specific accidents of curatorial history and the vagaries of changing cultural priorities over time more than they represent the actual facts of book production or consumption, even though -- paradoxically-- we have no better way of determining those facts.

In this respect, the proponents of "culturonomics" have perhaps something to learn from discourse analysis and the emphasis it places on the social and historical context within which speech data – whether written or spoken – occurs, is captured and curated. Context, the framing of an observation, the metadata around the data, is crucial no matter how big your big data may be, as Kath Bode amongst others has noted. In addressing the themes proposed for this seminar, I will first try to place the issues surrounding data and data modelling in a historical context, arguing that while it was once convenient to consider text as a special kind of data, today in a time of data-excess it is perhaps more appropriate to consider data as a special kind of text, with the same affordances and the same rugosities whatever form it may take.