# What does the textual scholar require of computer science?

Those whose work requires them to navigate the treacherous waters between documents and texts know well that the study of text is always an exercise in modelling. Computer science, except perhaps when it is purely about engineering, is above all about modelling and the manipulation of abstractions. This is why I consider its intellectual home to be within the Humanities, and why the Humanities have repeatedly appropriated its methods, even without always thoroughly understanding them.

## 1. Science and the humanities

A [recent diatribe by Roger Scruton](http://www.thenewatlantis.com/publications/scientism-in-the-arts-and-humanities) ('Scientism in the Arts and Humanities', in *New Atlantis*, Fall 2013) has got me thinking about that old chestnut “what is the digital humanities”. Scruton argues passionately and persuasively against what he terms “scientism” – the pretension to scientific method – in the humanities, reserving particular disdain for the notion of “research” in the humanities as the term is currently used by cross-disciplinary “xxxx-studies” in humanities departments across the English-speaking world. He points out that “research” in the sciences is concerned with the establishment by scientific method of evidence to support or refute a pre-existing hypothesis about the world, whereas in the Humanities, it is applied to just about kind of activity that may add to the sum of cultural knowledge at our collective disposal, or may simply act as a substitute for such knowledge. I was struck by his unashamedly Arnoldian appropriation of the term “culture” and what follows represents some further thoughts along similar lines.

If science aims to deepen our understanding of “the world”, and the humanities to deepen our understanding of “culture”, we do need to find a definition for culture which goes beyond simply saying (as Scruton does) that it is about the “I and I” (probably not so much a hint of Rastafarian influence as an insistence on the subjectivity of cultural thinking), though it is true that any account of culture which does not take into account its effects on the individual cultural consumer will be sadly lacking. The laws of physics operate whether we know about them or not; the same cannot be said of cultural norms. And yes, of course, culture, particularly “high culture”, is a social and political construct, reflecting or reacting against the social and political power structures of the context in which it is articulated, and thus seemingly entirely contextual and contingent. But this naïve cultural relativism simply ignores the effectiveness with which the very contingency of culture also reveals, often unconsciously, its context, enabling us to construct hypotheses around the social and political norms concerned, and to assess it with reference to a wider context. The pre-occupations of human culture have not changed so much over the centuries, though different reactions to (say) birth, sexual partnership, time, death, and the construction of society are readily discernible, as are different reactions to those reactions. It seems to me that a study of culture, in the sense for which the Germans used to use the term *geistwissenschaft ,* is a study of human reactions to, and constructions of, the world, and of those constructions. I further suggest that the relative merits of the various possible explanations it offers may be assessed in the same way as we evaluate purely scientific explanations.

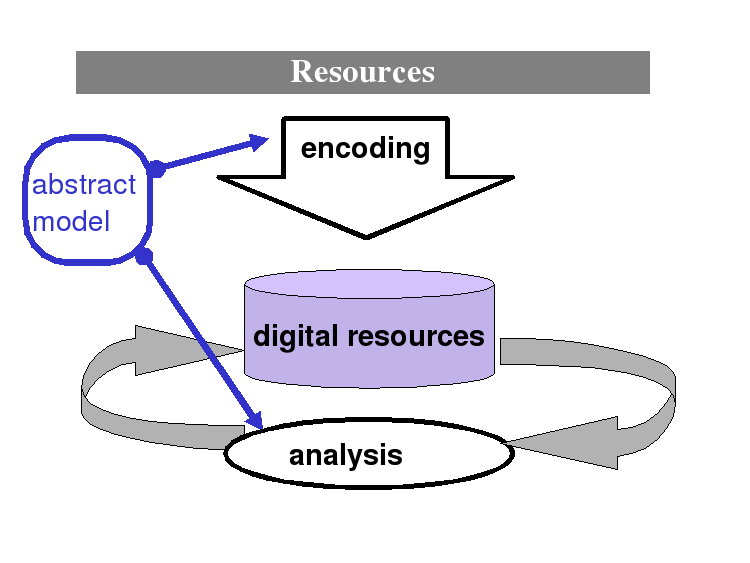
A scientific explanation is valued according to the effectiveness with which it provides evidence in support of a hypothesis. If however the hypothesis is very general, for example that there is a single elegant principle governing the behaviour of space, mass, and time, it may not readily be discernible as a hypothesis. When Eco says (in *Interpretation and overinterpretation*, 1992) that we value Copernicus' model of the Universe more than Ptolemy's not only because the former explains aspects left mysterious by the latter, but also because Copernicus enables us to understand the reasoning behind Ptolemy, he is not simply applying a humanistic perspective, exercising the hermeneutic meme to rhetorical effect, but demonstrating that evaluation should always proceed in the same way, whether we are considering the motions of the planets, or the relative merits of 19th century pulp fictions. For cultural objects do exist in the real world, and the cultural readings which confer “cultural” status upon them are also phenomena of the real world. Hence there is nothing inherently implausible about using scientific methods to gain some understanding of their behaviour, and how they function,

We should not however fall into the trap of supposing that in applying such methods to generate “scientistic” descriptions we have exhausted all there is of value in understanding a cultural object, a work of art. The history of cultural objects includes the history of its status considered as a work of art, but its meaning goes beyond the aggregation of perceptions of it as manifested in recorded opinions about it. Some of those perceptions may be ill-conceived or unhelpful, failing the Eco test of greater explanatory power for example, or other conceptual norms. To read *King Lear* solely as a political argument about kingship ignores the greater resonance of what it has to tell us about family life. To read *Hamlet* solely as an instance of the vogue for “revenge” tragedies that seems to have occurred on the English stage around the turn of the 16th century seems similarly wide of the mark. Contemporary African readers of Dickens' *Great Expectations* sometimes reduce it to a fable in which Pip's innocent life as a village dweller is corrupted by wealth and social class as soon as he moves to the town. Such a reading is one which Dickens might have recognised, and which the text certainly licences, but historically-minded critics may still feel that there is something wrong with implicitly equating the experience of a 20th century upwardly-mobile African villager and that of an imagined member of the 19th century rural poor. (Even so, a judgement we might anachronistically consider inappropriate when applied to a specific cultural product – for example, the use of racist or sexist terms in early 20th century writings – is surely appropriate when applied to the context in which such writings are created or delivered; indeed, the writings constitute essential evidence warranting such judgements.)

Consider, for example, linguistics. Language is surely the archetypal manifestation of a cultural object, almost a metaphor for culture itself (we talk about the “language of art”, we say that paintings and poets “talk to us” in a particular way, we even talk of a “vernacular” architecture). Over the last few decades, it has become increasingly clear that new technologies have facilitated a new perspective on the ways languages are used, hence how they change, and even perhaps what fundamentally they are. Corpus linguistics emphasizes the performative aspects of language, seeking to identify recurrent, possibly unconscious, regularities of usage, patterns which demand an explanation. Some have even claimed that no linguistic structures exist beyond such regularities of usage and the patterns associated with them, that there is no such thing as “grammar” analogous to the laws of physics in the real world. Even so, some explanation has to be found for these patternings. It is not necessary to subscribe to atavistic Chomskyan theories about innate grammar to seek explanations for them in terms of some general (and falsifiable) hypotheses about how languages function, for example to explain language variation and change by reference to a principle that innovation must always show itself first as deviation and is frequently associated with an assertion of group identity, that language users always value mutual comprehension above formal coherence or adherence to predefined norms, and so on.

So we should avoid claiming too much for a scientifically-derived and statistically-justified assessment of the facts of cultural reception. The history of digital humanities is dotted with the corpses of over-enthusiastic systematisers, from T.C. Mendenhall's “characteristic curves of composition” to J.A. Burrow's reduction of Jane Austen's style to vectors of frequency data, This not to say that statistical stylometry has nothing to tell us; just that it can only ever be a means to an end. The most scientific of stylometricians will always use the objective evidence revealed by their analysis in support of an entirely subjective judgment, be it about authorship or about style. As Stanley Fish did not quite say “There is *always* a text in this class”: in all such judgements, the constructed text, the reading of the evidence, is the end result of the research, whether it is obtained by meticulous statistical methods or good old fashioned introspection. And I tend to agree with Arnold, and with Scruton, that constructing such readings and transmitting them is actually the purpose of the Humanities.

A reading is however by no means the same kind of thing as a model. When I give introductory talks about the purpose and nature of text encoding, I often use the following schema to represent the distinction:

In its usual context, this schema is meant to show several things:

1. The process of transforming resources (cultural objects such as books, paintings, historical documents etc.) into digital form is always a form of re-presentation, abstraction, reduction, reinterpretation, or encoding. Or, one might say, reading.
2. The results of that transformation into digital form can be analysed and re-interpreted, automatically giving rise to an enriched version of that reading, which in turn can continue to be enriched in a kind of virtuous hermeneutic circle
3. The process of encoding, and the process/es of analysis, must however be informed by the same abstract model

Perhaps this is merely a long winded way of saying that you cannot get out of a system more than you put into it, but it does suggest that the conceptual model underlying a set of readings is a different kind of thing from any of those readings and operates at a higher level of abstraction.

## 2. Models in the Humanities and in Computer Science

I freely confess that my ideas about what the humanities are or should be were formed during a distant epoch: the end of the 1960s. And my ideas about what computer science is or should be were formed during one that feels even more remote : the end of the 1970s. I have also lived long enough in the hinterland between the two disciplines to see how the intellectual territory laid claim to by either has evolved.

In the 1960s, the discipline associated with the study of English literature, at least as I experienced it at Oxford University, was going through one of its periodic fits of self doubt. At other universities, these were tumultous times, as the waves of what was to become known as Theory (with a capital T) began to sweep away the Arnoldian consensus that works of art existed independently from their creators and consumers and were invested by an innate cultural value. Even at Oxford, traditionally skeptical about such French (or worse Cambridge) vulgarity as the search for theory, it was advisable to be aware of such theoreticians as Beardsley and Wimsatt, and to be able to shoot down the intentionalist and affective fallacies. We agreed that an understanding of its author's intention (as far as these could be determined) did not exhaust the meaning of a work of literature, any more than did an itemisation of its effects as recorded by its readers. We felt obscurely that we needed to place works within their historical and social context, to assess the extent to which they deviated from or confirmed reader expectation at different times, but we lacked the tools to do that, other than by the painstaking process of reading and remembering many, many books. We lacked an abstract model of how literature functioned or what it was, and did not know how to construct one.

The computer science I encountered at the end of the 1970s, by contrast, seemed obsessed by ways of representing knowledge, and constructing models. The entity-relationship model was succeeded by the Codasyl network model, which was blown away by the relational model, just as the giant mainframes began to be blown away by distributed networks of “personal” computers. Under the influence of large amounts of money and requirements for increasingly complex centralized information systems, these modelling techniques naturally evolved into methodologies such as SSADM (Structured Systems Analysis and Design Method, a set of standards developed in the early 1980s for systems analysis and [application](http://www.webopedia.com/TERM/A/application.html) design widely used for UK government computing projects) . It is easy to poke fun at that expansive pre-web era, in which modish re-brandings of essentially identical techniques succeeded each other with confusing regularity, always with extravagant claims of advanced capabilities just around the corner, in the “next generation” architecture. The next generation when it actually arrived in the nineties was distributed, decentralized, and almost entirely uninterested in all of the effort which the database designers and conceptual modellers of previous generations had put into trying to construct and impose a federated approach to the representation and storage of knowledge. (Which is why we now see the reinvention of logic programming in the form of linked data: but that is a different story.)

Nevertheless, like many others at the time, I found that the tools and techniques of computer science, though they might be described in terms of a particular jargon, and though their field of application might seem entirely alien, still had something to offer the humanist. Could it be that an abstract model for the way that texts and documents function – which I take to be the essential business of the humanities – might be expressed using the same language as that used to model the data flows and processing requirements of East Midlands Gas?

It seemed clear to me that texts and documents should be described from at least three perspectives [an idea I first articulated in Burnard 1987]: as physical objects with a visual representation; as linguistic objects made up of words and phrases drawn from some kind of linguistic system; and as intensional objects with reference to real world objects, events, or entities. Most computer systems of the time tended to prefer one or other of these aspects. A word processor would help you produce nice printed copies of your documents; an information retrieval system would help you investigate their language; a database would help you describe what they were about. Systems which crossed these frontiers, enabling you to control the appearance of particular words used to describe felonious transactions in a court record, for example, were harder to find, and usually required to be custom-built, with many compromises along the way.

With the arrival of markup languages such as SGML in 1987 and XML a few years later, it became possible at last to describe a document in a detailed way independently of which of these three aspects was to predominate in its processing, and hence in a way that facilitated all of them equally. And with the arrival of the Text Encoding Initiative around the same time, an extraordinary adventure in document modelling got underway. Much has been written about the TEI (not all of it by me) and its significance; my favourite comment is that whatever else we may say of the TEI *Guidelines,* as Basil Bunting did of Pound’s ‘*Cantos’,* “they resemble the Himalayas: you can ignore them if you like – but you will have to go an awfully long way round.” The TEI's relevance to the present paper is that it represented the first and so far only time that scholars from across the humanities disciplines were succesfully corralled into achieving some kind of consensus about the “significant particularities” of the documents they studied. The TEI was (and perhaps remains) a unique exercise in inventorising the components of the models underlying research in the humanities, from the disparate points of view of lexicographers, linguists, critical editors, manuscript scholars, historians, literary scholars, and librarians. To find an abstract language adequate to represent such divergent perspectives within a single framework we naturally sought to apply data modelling techniques inherited from computer science, expressed not in UML or SQL but using the new text-friendly features of SGML. The rest, as they say, is history.

If the success of the TEI shows us that the modellimg techniques inherent to computer science could successfully be imported and made to function within the humanities paradigm, it seems reasonable to conclude by asking whether this is a unique instance of such synergy.

## 3. What did the computer scientists ever do for us?

Unsurprisingly perhaps, the things that working textual scholars seem to most appreciate about the impact of information technologies on their working practices are all things that computer science as a discipline tends to take for granted. When asked a version of my working title for this piece, an English professor of my acquaintance replied:

“I think I owe the discipline a great deal... the advantages of on-line ordering before a visit to the British Library (say), and (from home) easy access to bibliographical and biographical information when preparing a book or essay ms for the press. I've regularly used Google to track down unattributed quotations which might otherwise have taken me ages to locate; I also use the electronic databases ECCO and EEBO, although I think the interfaces and general tractability have some way to go. I ought to add, I think, the sheer convenience of being able to assemble large and complex texts--such as editions—electronically, where relevant information comes to hand over a period of time. I only wish word-processing had been available when I completed my Ph.D. in 1979... Finally, I assume that without information science there could be no email, and without email I think that academic exchange as we know it might grind to a halt.”

This reply perhaps demonstrates how deeply embedded information science has become. The aspects selected by my colleague – networked access to information resources both of the kind traditionally held by libraries and of the kind traditionally embodied in one's peers – constitute a change in the knowledge infrastructure, the context in which work is done. There is much to be added if we are to give an adequate account of that infrastructure: about the politics of open source access, about the alleged democratising (or to use the French word *vulgarisation*) of access to cultural resources, about all the ways in which the Internet has transformed our ways of knowing about of the world, and the world that we know about. “Never before have so many people known so little about so much” … but the changes are now driven by commercial and social imperatives more than they are by the interplay of academic disciplines which is my subject.

My colleague's reference to word processing also hints at a more subtle change in the way that work is done itself. Of course writing on a word processor is only superficially like writing on a typewriter, just as a typewriter is only superficially like a quill pen. The extent of quantitative change in going from a machine in which making corrections is an expensive and limited process to one in which documents are never finished such is their fluidity and plasticity really does approximate to a qualitative change. In the 90s this occasioned anxiety about apparently fundamental shifts in the very nature of scholarly communication, even the thinking process itself, induced by the spread of new technologies. A couple of decades later, in a seemingly entirely fragmented, and decentred world, drowning in media which seem to be dominated by twitter and sound bites, we do well to remember the positive side of that transformation.

In placing first the availability of digital resources, however, I think my colleague hits the mark exactly. The challenge for computer science has always been to find better tools for coming to terms with information glut, whether in the form of paper archives or millions of digitized books. The success of Google may have suggested to some that the indexing and cataloguing techniques associated with classical information retrieval techniques were entirely superannuated. But that model of the document as witness to a mode of expression, a particular discourse, suggests that such a view is premature. Indexing techniques are beginning to take on new and more sophisticated clothing, their function rebranded as text mining or text modelling. If it is the words that conspire to form the meaning of a text, we should be able to able to formulate new, more coherent, and better informed hypotheses about that meaning on the basis of their relative co-occurrences and absences in the immense bodies of digital text now at our disposal. To quote Ted Underwood, “The notion that documents are produced by discourses rather than authors is alien to common sense, but not alien to literary theory.” As we do so, the availability for the first time of massive quantities of digital text structured and organized in terms of our traditional models of text and textuality (rather than their purely visual properties) will enable us to make richer (and thus more explanatory) models against which to judge the salience of individual works, and in terms of which to categorise their context. Rather than looking for the proverbial needle in a haystack, we should start considering why hay is such a good home for it. I do not know whether that is a notion that computer science has fully assimilated as yet.