Fact—Information—Data—Knowledge: Databases as a Way of Organizing Knowledge

LYNETTE HUNTER
University of Leeds

Abstract

Many workers in the area of language and literature assume that users of computers are seeking to quantify aesthetic 'results'. However, there is nothing in computer use which necessarily leads to the decontextualised or reductive logic of quantification. Indeed the use of computers can sharply foreground those contexts and contribute to the construction of valid actions and assessments of action that the humanities pursue. It has to be said, though, that some computer users are convinced by their colleagues and do pursue reductive activities. This also results from the historical cohabitation of modern science with mathematics, and the reductive paths of its technical application which transpose onto computer use in general.

To meet the challenge of current institutional relegation, computer users in the humanities need thoroughly to investigate the grounds of their methodologies. Just as the social sciences in the 1960s and 70s, whose experience is at least partly analogous to our own, we need to explore the common grounds of our often very specialised activities and to study the implications of our responses so far to the strategies we have been offered. The discussion attempts to locate some common ground by introducing the ways in which all people in the humanities use a strategy which is held to be fundamental to computer use, the organization of data. It also attempts to extend our understanding of that common ground by considering the organization of data as one of the underlining methodological features of all computer applications in the humanities, and by suggesting a number of paths that could be explored in any attempt to define the broader ideological implications of what we are doing.

Introduction

There is always a difficulty in determining an audience when discussing computer use in the humanities' area. The majority of people interested in participating in such a discussion may well have personal computers, and may even use them for word-processing. A smaller group will be quite experienced, probably with considerable expertise in one or more fields of specialised application such as text preparation or concordancing or bibliography. An even smaller group is made up of those who do not have computers but who wonder whether they might be useful. Beyond these potential participators is the vast number of people who actively reject or evade the recognition of computer use. They do so first because they are under the misapprehension that computers engage with a radically different epistemology to the one the humanities value and promote, and second because there is a residual fear that computer use will negate and undermine the ideological assumptions of current cultural enquiry (Hunter, 1990a).

Correspondence: Lynette Hunter, School of English, University of Leeds, Leeds LS2 9JT, UK. E-mail: ENG6LH@UK.AC.LEEDS.CMS1

However, computer use is an epistemological strategy that is neither inherently disruptive of conventional value nor helpful to it. Rejection of it closes the door on an opportunity to question the way in which we structure the world. Furthermore this rejection encourages at one and the same time a hermetically sealed group of computer users and institutional pressures on those users to come up with exactly the procedures that negate the methodologies of humanities students and scholars. Many workers in the area of language and literature assume that users of computers are seeking to quantify aesthetic 'results': they relegate them to a kind of maths club for second-rate academics who could not prove their moral worth in the accepted activities of topical and contextual reasoning that have defined the humanities from their initial construction in the policies of humanist educationists in the Renaissance period (Hunter, 1990b). Not, of course, that this relegation is consciously articulated. Those willing to engage in such articulation would also be aware of the main fault in the argument, which is: that there is nothing in computer use which necessarily leads to the decontextualized or reductive logic of quantification. Indeed the use of computers can sharply foreground those contexts and contribute to the construction of valid actions and assessments of action that the humanities pursue. It has to be said, though, that some computer users are convinced by their colleagues and do pursue reductive activities. This also results from the historical cohabitation of modern science with mathematics, and the reductive paths of its technical application which transpose on to computer use in general (Hunter, 1991).

To meet the challenge of current institutional relegation, computer users in the humanities need thoroughly to investigate the grounds of their methodologies. Just as the social sciences in the 1960s and 70s, whose experience is at least partly analogous to our own, we need to explore the common grounds of our often very specialized activities and to study the implications of our responses so far to the strategies we have been offered. The discussion that follows is addressed both to those who use and those who do not use computers. People working within literary and linguistic computing will be familiar with many of the activities described, and it worth noting that these activities are orientated toward the analysis, criticism and commentary on verbal texts rather than toward teaching tools.[1] What this discussion attempts to do is to recast those activities in terms of the way they order and structure our observations of verbal texts. The discussion attempts to locate some common ground by introducing the ways in which all people in the humanities use a strategy which is held to be fundamental to computer use, the organization of data.

organize observation, and the need constantly to remind ourselves of what has been omitted. The reason that numbers are so powerful is that they arise out of a set of definable human premises or rules or axioms, and as long as they stay within these they can give 'true' answers. But people and language don't work like that. Numbers often find it helpful to forget what is omitted, but language and literature fail if they do so. The problem many humanities users fear, and which is evidently present in the responses to databases and computers in the study of language and literature, is to do with turning language and literature into reductive fact and forgetting not only the conceptual procedures of the process but also the social context. There is an enormous amount of philosophical and epistemological work to be done in the area of what 'data' is for the arts, and what kind of knowledge people involved in computing for the arts are pursuing.

If observations become data once we start questioning them because we impose a structure through our questions, then data becomes the key to the concept of significance. Significance is partly to do with the recognition of patterns and structures that our questions both desire and define. Two primary procedures for that recognition are hypothesis, which deals with the testable, and story or narrative, which deals with the contextual. Broadly speaking these two procedures outline the approaches of, respectively, the sciences and the humanities. Hypothesis is based on a logic that aims toward the validating of possible worlds that are consistent in their own terms. On the other hand narrative is structured through topical reasoning that enacts the perceived interactions of current contexts and is often inconsistent. In coming to terms with computer use in the humanities, we need to understand both that hypothesis is an unreflective version of topical analogy, and that such selective masking of context can be helpful in its own ways.

Verbal texts as ordered structures

Disregarding the possibility that the questions themselves may be the organizing agents, a database may be taken to be any set of observations ordered or structured in such a way as to answer certain questions. In the study of language and literature, a text or reported speech becomes recognisable as data the moment we ask questions of it that it can answer. One way of approaching the use of computer databases in the field might be to look at the kinds of questions we already ask and in what sense we already treat texts and speech as databases. The textual critic asks a wide range of questions of the linguistic text at hand: questions about narrative, character, voice, verse, prosody, language, genre, persuasion and so on. The questions are usually asked because we are interested in the topic, in context, in social and historical relations. Studies of vocabulary, figures, structure, all underlie studies of context and topic—although they can have no significance divorced from them.

Frequently the reader is looking for a pattern in the text, even if it is a pattern that disrupts or is disrupted. To recognize a pattern the reader or critic will look for discernible elements which are taught to us as part of reading and writing skills: an understanding of lexis, of grammar, of communication, or rhetoric. Theories in literature and language criticism often arise as attempts to suggest ways of looking for different elements: for example the now rather narrowing pairing by Jakobson of metaphor and metonymy, or Genette's narratology, or Booth's irony, or Halliday and Hasan's cohesion (Jakobson, 1956, Genette, 1966, Booth, 1974, Halliday and Hasan, 1976). Anyone engaged in studying verbal texts knows that the kind of questions asked are immediate to the individual: we all ask differently even if within ideological parameters. Recognising a pattern implies remaining open to gatherings, groupings, clusters, repetitions, and responding to the internal and external relations they set up.

Let's say we're doing a good old-fashioned practical criticism exercise on the openings of two novels: among the elements with which we are concerned will be how the voice speaking to us defines itself, what is its relation with the reader? With the characters? In order to talk about the elements of narrator, perspective or point of view, tone (irony, satire, etc), stance (rhetorical invitation)—which we need to do in order to assess the contextual implications—we will need to take a number of strategies out of the critical history and go looking for any or all of a variety of things such as adjectives, forms of address, repetition of words, use of factual detail such as proper and place names, and dates, and so on. We may need dictionaries, thesauri, books of quotations and aphorisms, historical sources, and many other reference-points. Each time we focus on a specific element we are foregrounding it as significant to our enquiry; we are suggesting that it will answer our questions; we are in effect imposing an organizing structure on it and turning it into data.

When critics carry out even a simple piece of study such as this, they often work according to an agreed consensus specific to a particular critical context. The New Critics worked in one way and the structuralists in another. Furthermore, the work is done with a greater or lesser emphasis on outlining the methodology supported by that consensus. We have probably all had experience of critics who expect us to take for granted the value of studying some element of which we no longer can see the use, say naive authorial intention; and conversely the experience of critics who expect us to accept without question highly sophisticated analyses in a vocabulary they do not explain. In either case, and in many others, there has been a profound lack of self-reflexivity in many discourses of the humanities, and recognition of this has led to the recent upsurge in courses on critical theory and cultural studies. But the more we become aware of the bases for our questions the more we will be able to assess what 'data' is for our discipline, and then perhaps we may be able to assess the sense or nonsense in the use of computerized databases for a particular piece of work.

Computerized editions or transcripts of verbal texts

Given that when pushed most critical readers can point to elements that they take to be significant and explain

choices and presentation, so the selection of certain elements in a computerized text as significant will be historically specific to the needs of readers. The selection of significant elements is the key to database construction, and in a sense this discussion has jumped ahead chronologically in order to make clear the analogies between normal literary critical activity and that assisted by the computer.

Computers are only machines, and early computers could only carry out fairly crude operations on relatively short texts. Many of the early databases constructed for literary and linguistic study selected significant elements right out of the text and placed them in pre-determined fields to make it possible for the computer to perform matching and comparing operations. Most of the early work in this field was done in authorship studies (Morton, 1978). Its often naive approaches may be recuperated as constructive both through the social context they can make relevant when for example you are suddenly able historically to locate a piece of writing, and through the internal evidence of the use of vocabulary and syntax specific to a particular voice in a general historical setting.[11] Partly because of the difficulty of studying elements abstracted from textual context without a nervous worry about reductive narrowing, and partly because for centuries work in bibliography had been trying to restore a non-verbal context to literature and language by means of overtly organizing techniques analogous to those employed by computer strategies, many other early computer-assisted studies relevant to verbal texts were carried out by bibliographers interested in the social and historical context addressed by the printed medium for communicating words.

The classic database in bibliography is the card index of Author, Title, Publisher, Date; and as the basis for early computer databases these were called 'flat file' databases.[12] They have the same type of information on each card, and the same fields for each entry. The good thing about them is that given that you do not want to change the distinct fields of information, they can be adapted for a number of different database questions, or programs, relatively easily, because you have done a lot of work already by selecting and defining the significant material yourself. They also, for this reason, work quite fast in computer terms: essentially the program, or set of possible questions that you choose, runs around matching up previously defined fields. The main thing that makes one program different from another is the way in which it can match up different fields.

An example of this kind of database, and there are many, would be one that a group of researchers (Attar, Driver, Hunter) created of domestic texts published in Britain between 1800 and 1914.[13] There are around 3000 separate individual texts, and roughly 25,000 distinctly different editions of the texts, so 25,000 different index cards would be needed. This represents a substantial amount of description. Further, each card contains fields for Author, Title, Publisher, Collation, Description, Citation, Location, Notes, Topic, and more than 40 subfields. The database was initially set up for the FAMULUS (Burnard, 1985) database structure, which allows for swift indexing and which was particularly helpful for generating chronological, short-title, and topic indexes. Because so much work had gone into marking up different kinds of potentially significant detail, the database was also adaptable to other structures such as STATUS or EXTRACT[14] which allow questions and combinations of questions about for example which authors were publishing with a specific publisher between 1860 and 1870? or what on average did domestic books published in London during the 1890s cost compared to those published in the regions?

This type of selective database is a powerful research tool and several, such as the Eighteenth Century Short-Title Catalogue or the Nineteenth Century Project (Crump, 1989), are available. What they offer impinges directly as questions of genre, audience, authorship, copyright, censorship, and printing history. Provided that care is taken with the decision about how to organize information, this type of 'flat-file' database can be relatively flexible and very helpful not only to bibliographers and literary critics, but also to historians of publishing, society, economics, and culture. Its limitations lie in the historically specific necessity of selection for significance. In a sense it leaves itself even more open to criticism of its selection procedures than a computerized textual edition because it omits so much; at the same time simply because of years of bibliographic tradition, there is more agreement on the kinds of description appropriate to the audience the bibliography is trying to reach.

A selective database is a particular kind of written text, often implicitly a piece of criticism or historical commentary structured in such a way as to address the practical problems of organizing response to a large area of material, and to encourage the audience to respond by reorganizing the descriptions so that they can answer different questions. Once a crossword has been completed there is not much more to be said; it provides a severely defined mode of organization. Yet Propp's observations on literary motifs in fairy tales have been re-organized by linguists, anthropologists, narratologists and many others (Eagleton, 1989). However, we are not particularly sophisticated in structuring these texts; we have a lot to learn. While literary texts can depend upon a fairly stable consensus with regard to vocabulary, grammar, rhetoric, and poetic, selective databases need to address the problem of longevity more urgently. The needs of future audiences have to be considered and guessed at; and one way of extending their useful life is to make them in such a way that people in 50 years' time can reorganize the observations and create structures more appropriate for their needs. For example in the case of the bibliography of domestic texts, although the average reader might only want to know who published a book, it could be conjectured that at some time in the future someone might want to be able to separate out the different names in the publishing imprint in order to follow associations and co-operative activity between different publishing firms during a particular period. To facilitate such a question some kind of mark-up can be introduced into the initial construction of the computerized text. Of course it is impossible to speculate on all the future questions, but some consideration of them is invaluable in the long-term. It is at this level that the user may begin to wonder whether the computer couldn't be trained to do some of this for us.

curved is an apogee of something. We are also, in our society, relatively sophisticated at reading maps, so that when A. Morton maps collocations in Shakespeare authorship studies on top of each other while T. Merriam[18] maps them along a binomial distribution in discreetly presented visual shapes, we think the Merriam more persuasive, more convincing. Somehow when Burrows produces through Eigenvalue calculations individual 'voiceprints' for a writer which shift chronologically but appear to retain unique shape, we believe in this while we simultaneously lecture to our students on the disappearance of the essential identity of the author (Erickson and Nosanchuk, 1983). Possibly we bring to this our insistent exposure to the cultural use of geometric forms that has also characterized our culture since the Renaissance.

There is, as with all modes of ordering, nothing intrinsically wrong with statistics. Its organization, like that of the bibliographic database, is simply further removed from the text and hence omits more than a computerized edition. If you like, statistics itself is a kind of second- or third-order database. But two points should be made: if humanities scholars or students are to proceed more often into the possibilities for numerical analysis offered by the sheer quantity and detail of computerized databases, then both those using these techniques and those attempting to assess them should understand far more about them. Not much attention has been paid to what the literary or linguistic critic does with either numerically- or visually-indicated significance. Contributions to the discussion that took place on these issues during the 1970s in the social sciences have emphasized the difference between exploratory and confirmatory statistics. Explanatory statistics are those which guide you to a number of potential hypotheses, while confirmatory statistics test these hypotheses. The distinction is helpful, but it is unfortunate that the usual scientific environment for 'hypothesis' uses it to establish testable validity. Just so, there is a tendency for humanities' users to employ statistics to imply that their results are in a sense 'true' because testable, and for those who don't use statistics to reject them as spurious because they eliminate and omit so much context. At the moment it would suggest that the responsibility lies with those who use them to indicate the methodology and hence provide the context, if they want their studies to be helpful.

Possibly, for critics of literature and language, the recent developments in hypertext databases are more immediately fruitful because they emulate the way we already research and are therefore circumscribed by well-tried guidelines that we currently agree to be acceptable. This is hypertext's most attractive but also most limiting feature. As yet there is no such thing as a typical hypertext database, but two models quickly emerge. The first centres the database around a specific text, say one of Austen's novels again, and adds many other recognized research tools to it such as a thesaurus, a dictionary, critical works, social and historical background text, encyclopaedias, and even other programs such as concordancing packages. The central novel can be marked-up or can remain visually without markup but responsive to a scan by cursor. In either case, words or phrases or sections can be indicated to signify points at which the person constructing the database thinks a link to a definition, a reference, a comment, or another text might be valuable. Different hypertext organisations will work in rather different ways. The Macintosh HYPERCARD allows the organiser essentially to create a flat file database with initial text rather than selective descriptions on each card, which texts can be related hierarchially or serially to each other.[19] GUIDE allows the organizer to create a database containing many different texts among which the user can wander.[20]

The key to hypertext organization is that it allows us to digress, to add narrative at significant points of the text. In a manner parallel to computerized textual editing, nothing need be omitted from the text but signifcant elements can be signalled. But whereas the editing of the computerized text signals elements to be drawn out of the language or literature, to be located, collocated, counted and analysed, the editing of a hypertext text signals elements that could be extended, explained, related or commented upon. Instead of potentially abstracting from context the hypertext can insist on more context within the database itself. The insistence is double-edged for while on the one hand the user of the hypertext database will find it difficult to avoid context, that user will also be encouraged to read within the context provided. The very friendliness of a hypertext leads to its drawback, which is that it is much more difficult to foreground the kind of guidance, the method for indicating significance, in the organization of the included material. Because it can be so close to our expected academic procedures, it is more difficult to foreground.

The second model for hypertext database organization, which is possibly more overtly reflexive, considers the database as a group of interconnecting texts. The Jane Austen database would then consider the novel, the critical texts, the dictionary and so on, as equally important contexts which it could be helpful to relate to each other. While this can be interesting because it begins to challenge our concepts of the writer's subjective authority, it is also extremely difficult to ensure that the user uses it in this way. No doubt guidelines, rules of thumb, and many different models, will emerge in the hypertext media. However, at the moment, they do appear to encourage organization of textual descriptions into non-numerical patterns based on topical reasoning and predicated on social and historical context. They encourage us to tell stories rather than generate statistically-testable hypotheses. I would not want to suggest that hypertext structure can answer more questions for the humanities' user than those of other database structures, but they do offer the possibility of foregrounding the narrative strategies we conventionally use in the practice of criticism and hence also pose questions about our methodology in a more immediately challenging and recognisable way.

The most immediate question that is raised concerns what kind of significance we are looking for. If we look for similarities, variants, comparison, repetitions, the implicit base for our analysis of pattern and structure is numerical and can allow us to generate hypotheses. We may also, in contrast, go looking for patterns based on

Hockey, S. and Martin, J. (1988). *Oxford Concordance Program*. Oxford: Oxford University Computing Service.
Hunter, L. (1984). *Rhetorical Stance in Modern Literature*. London: Macmillan.
Hunter, L. (1989). *Modern Allegory and Fantasy*. London: Macmillan.
Hunter, L. (1990a). 'The Computer as Machine: Friend or Foe?'. In C. R. R. Turk (ed.), *Humanities Computing*, London: Kogan Page.
Hunter, L. (1990b). 'From Cliche to Archetype'. In L. Hunter (ed.), *Toward an Understanding of Analogical Reasoning*. London: Macmillan.
Hunter, L. (1991). 'Rhetoric and Artificial Intelligence'. In R. Roberts (ed.), *Rhetoric and the History of the Human Sciences*. Bristol: Bristol Classics Press.
Jakobson, R. (1956). 'Two Aspects of Language and Two Types of Aphasic disturbances'. In R. Jakobson and M. Halle (eds.), *Fundamentals of Language*. The Hague: Mouton.
Morton, A. (1978). *Literary Detection*. New York: Scribner.
O'Shea, T. (1987) 'Machine Learning'. In T. O'Shea, J. Self and G. Thomas (eds.), *Intelligent Knowledge-Based Systems*. London: Harper and Row.
Stigler, S. (1986). *The History of Statistics*. London: Belnap Press.
Urkowitz, S. (1980). *Shakespeare's Revision of King Lear*. Princeton: Princeton University Press.