1.

The History of Humanities Computing

Susan Hockey

Introduction

Tracing the history of any interdisciplinary academic area of activity raises a number of basic questions. What should be the scope of the area? Is there overlap with related areas, which has impacted on the development of the activity? What has been the impact on other, perhaps more traditional, disciplines? Does a straightforward chronological account do justice to the development of the activity? Might there be digressions from this, which could lead us into hitherto unexplored avenues? Each of these questions could form the basis of an essay in itself but within the space and context available here, the approach taken is to present a chronological account which traces the development of humanities computing. Within this, the emphasis is on highlighting landmarks where significant intellectual progress has been made or where work done within humanities computing has been adopted, developed or drawn on substantially within other disciplines.

It is not the place of this essay to define what is meant by humanities computing. The range of topics within this *Companion* indeed sends plenty of signals about this. Suffice it to say that we are concerned with the applications of computing to research and teaching within subjects that are loosely defined as "the humanities", or in British English "the arts." Applications involving textual sources have taken center stage within the development of humanities computing as defined by its major publications and thus it is inevitable that this essay concentrates on this area. Nor is it the place here to attempt to define "interdisciplinarity" but by its very nature, humanities computing has had to embrace "the two cultures", to bring the rigor and systematic unambiguous procedural methodologies characteristic of the sciences to address problems within the humanities that had hitherto been most often treated in a serendipitous fashion.

Beginnings: 1949 to early 1970s

Unlike many other interdisciplinary experiments, humanities computing has a very well-known beginning. In 1949, an Italian Jesuit priest, Father Roberto Busa, began what even to this day is a monumental task: to make an *index verborum* of all the words in the works of St Thomas Aquinas and related authors, totaling some 11 million words of medieval Latin. Father Busa imagined that a machine might be able to help him, and, having heard of computers, went to visit Thomas J. Watson at IBM in the United States in search of support (Busa 1980). Some assistance was forthcoming and Busa began his work. The entire texts were gradually transferred to punched cards and a concordance program written for the project. The intention was to produce printed volumes, of which the first was published in 1974 (Busa 1974).

A purely mechanical concordance program, where words are alphabetized according to their graphic forms (sequences of letters), could have produced a result in much less time, but Busa would not be satisfied with this. He wanted to produce a "lemmatized" concordance where words are listed under their dictionary headings, not under their simple forms. His team attempted to write some computer software to deal with this and, eventually, the lemmatization of all 11 million words was completed in a semiautomatic way with human beings dealing with word forms that the program could not handle. Busa set very high standards for his work. His volumes are elegantly typeset and he would not compromise on any levels of scholarship in order to get the work done faster. He has continued to have a profound influence on humanities computing, with a vision and imagination that reach beyond the horizons of many of the current generation of practitioners who have been brought up with the Internet. A CD-

ROM of the Aquinas material appeared in 1992 that incorporated some hypertextual features *("cum hypertextibus")* (Busa 1992) and was accompanied by a user guide in Latin, English, and Italian. Father Busa himself was the first recipient of the Busa award in recognition of outstanding achievements in the application of information technology to humanistic research, and in his award lecture in Debrecen, Hungary, in 1998 he reflected on the potential of the World Wide Web to deliver multimedia scholarly material accompanied by sophisticated analysis tools (Busa 1999).

By the 1960s, other researchers had begun to see the benefits of working with concordances. A series of four articles by Dolores Burton in the journal *Computers and the Humanities* in 1981–2 attempted to bring these together, beginning with a discussion of the 1950s (Burton 1981a, 1981b, 1981c, 1982). Some of these researchers were individual scholars whose interests concentrated on one set of texts or authors. In the UK, Roy Wisbey produced a series of indexes to Early Middle High German texts (Wisbey 1963). In the USA Stephen Parrish's concordances to the poems of Matthew Arnold and W B. Yeats introduced the series of concordances published by Cornell University Press (Parrish 1962). This period also saw the establishment of computing facilities in some major language academies in Europe, principally to assist with the compilation of dictionaries. Examples include the Trésor de la Langue Française (Gorcy 1983), which was established in Nancy to build up an archive of French literary material, and the Institute of Dutch Lexicology in Leiden (De Tollenaere 1973).

Although much activity at this time was concentrated on the production of concordances as ends in themselves, one application of these tools began to take on a life of its own. The use of quantitative approaches to style and authorship studies predates computing. For example, Augustus de Morgan in a letter written in 1851 proposed a quantitative study of vocabulary as a means of investigating the authorship of the Pauline Epistles (Lord 1958) and T. C. Mendenhall, writing at the end of the nineteenth century, described his counting machine, whereby two ladies computed the number of words of two letters, three, and so on in Shakespeare, Marlowe, Bacon, and many other authors in an attempt to determine who wrote Shakespeare (Mendenhall 1901). But the advent of computers made it possible to record word frequencies in much greater numbers and much more accurately than any human being can. In 1963, a Scottish clergyman, Andrew Morton, published an article in a British newspaper claiming that, according to the computer, St Paul only wrote four of his epistles. Morton based his claim on word counts of common words in the Greek text, plus some elementary statistics. He continued to examine a variety of Greek texts producing more papers and books concentrating on an examination of the frequencies of common words (usually particles) and also on sentence lengths, although it can be argued that the punctuation which identifies sentences was added to the Greek texts by modern editors (Morton 1965; Morton and Winspear 1971).

It is believed that the first use of computers in a disputed authorship study was carried out on the Junius Letters by Alvar Ellegard. Published in 1962, this study did not use a computer to make the word counts, but did use machine calculations which helped Ellegard get an overall picture of the vocabulary from hand counts (Ellegard 1962). What is probably the most influential computer-based authorship investigation was also carried out in the early 1960s. This was the study by Mosteller and Wallace of the *Federalist Papers* in an attempt to identify the authorship of the twelve disputed papers (Mosteller and Wallace 1964). With so much material by both authorship candidates on the same subject matter as the disputed papers, this study presented an ideal situation for comparative work. Mosteller and Wallace were primarily interested in the statistical methods they employed, but they were able to show that Madison was very likely to have been the author of the disputed papers. Their conclusions generally have been accepted, to the extent that the *Federalist Papers* have been used as a test for new methods of authorship discrimination (Holmes and Forsyth 1995; Tweedie et al. 1996).

At this time much attention was paid to the limitations of the technology. Data to be analyzed were either texts or numbers. They were input laboriously by hand either on punched cards, with each card holding up to eighty characters or one line of text (uppercase letters only), or on paper tape, where lower-case letters were perhaps possible but which could not be read in any way at all by a human being. Father Busa

has stories of truckloads of punched cards being transported from one center to another in Italy. All computing was carried out as batch processing, where the user could not see the results at all until printout appeared when the job had run. Character-set representation was soon recognized as a substantial problem and one that has only just begun to be solved now with the advent of Unicode, although not for every kind of humanities material. Various methods were devised to represent upper- and lower-case letters on punched cards, most often by inserting an asterisk or similar character before a true upper-case letter. Accents and other non-standard characters had to be treated in a similar way and non-Roman alphabets were represented entirely in transliteration.

Most large-scale datasets were stored on magnetic tape, which can only be processed serially. It took about four minutes for a full-size tape to wind from one end to the other and so software was designed to minimize the amount of tape movement. Random access to data such as happens on a disk was not possible. Data had therefore to be stored in a serial fashion. This was not so problematic for textual data, but for historical material it could mean the simplification of data, which represented several aspects of one object (forming several tables in relational database technology), into a single linear stream. This in itself was enough to deter historians from embarking on computer-based projects.

Representation problems extended far beyond specific characters. For concordance and retrieval programs there was a need to identify citations by their location within the text. The methods used by conventional document retrieval systems were inadequate because they tended to assume document structures similar to those of journal articles and were unable to cope with the structures found in poetry or drama, or in manuscript sources where the lineation is important. Various methods of defining document structures were proposed, but the most sophisticated one developed at this time was that used by the COCOA concordance program (Russell 1967). Modeled on a format developed by Paul Bratley for an Archive of Older Scottish texts (Hamilton-Smith 1971), COCOA enables the user to define a specification for the document structure which matches the particular set of documents. It also enables the markup of overlapping structures, making it possible, for example, to encode a citation system for a printed version in parallel with that for the manuscript source of the material. COCOA is also economical of file space, but is perhaps less readable for the human.

The other widely used citation scheme was more dependent on punched card format. In this scheme, often called "fixed format", every line began with a coded sequence of characters giving citation information. Each unit within the citation was positioned in specific columns across the line, for example the title in columns 1–3, verse number in columns 5–6, and line number in columns 7–9. The entry of this information was speeded up by functions on the punched card machine, but the information also occupied more space within the computer file.

The legacy of these citation schemes can still be found in electronic texts created some time ago. COCOA, particularly, was very influential and other schemes were derived from it. COCOA cannot easily handle the markup of small features within the content such as names, dates, and abbreviations, but its ability to deal with overlapping structures outstrips that of almost all modern markup schemes.

This period also saw the first opportunities for those interested in humanities computing to get together to share ideas and problems. In 1964, IBM organized a conference at Yorktown Heights. The subsequent publication, *Literary Data Processing Conference Proceedings*, edited by Jess Bessinger and Stephen Parrish (1965), almost reads like something from twenty or so years later, except for the reliance on punched cards for input. Papers discuss complex questions in encoding manuscript material and also in automated sorting for concordances where both variant spellings and the lack of lemmatization are noted as serious impediments.

As far as can be ascertained, the Yorktown Heights conference was a one-off event. The first of a regular series of conferences on literary and linguistic computing and the precursor of what became the Association for Literary and Linguistic

Computing/Association for Computers and the Humanities (ALLC/ACH) conferences was organized by Roy Wisbey and Michael Farringdon at the University of Cambridge in March, 1970. This was a truly international event with good representation from both sides of the Atlantic as well as from Australia. The proceedings, meticulously edited by Wisbey (1971), set the standard for subsequent publications. A glance through them indicates the emphasis of interest on input, output, and programming as well as lexicography, textual editing, language teaching, and stylistics. Even at this time the need for a methodology for archiving and maintaining electronic texts was fully recognized.

Another indication of an embryonic subject area is the founding of a new journal. *Computers and the Humanities* began publication in 1966 under the editorship of Joseph Raben. With characteristic energy, Raben nurtured the new journal and during its first years, at least until the regular series of conferences and associations that developed from them got going, it became the main vehicle for dissemination of information about humanities computing. Raben recognized the need just to know what is going on and the journal's Directory of Scholars Active was the first point of call for people who were thinking about starting a project. Other informal newsletters also served specific communities, notably *Calculi* for computers and classics, edited by Stephen Waite.

The 1960s also saw the establishment of some centers dedicated to the use of computers in the humanities. Wisbey founded the Centre for Literary and Linguistic Computing in Cambridge in 1963 as support for his work with Early Middle High German Texts. In Tübingen, Wilhelm Ott established a group which began to develop the suite of programs for text analysis, particularly for the production of critical editions. The TuStep software modules are in use to this day and set very high standards of scholarship in dealing with all phases from data entry and collation to the production of complex print volumes.

Work in this early period is often characterized as being hampered by technology, where technology is taken to mean character sets, input/output devices and the slow turnaround of batch processing systems. However, researchers did find ways of dealing with some of these problems, albeit in a cumbersome way. What is more characteristic is that key problems which they identified are still with us, notably the need to look at "words" beyond the level of the graphic string, and to deal effectively with variant spellings, multiple manuscripts, and lemmatization.

Consolidation: 1970s to mid-1980s

If any single-word term can be used to describe this period, it would almost certainly be "consolidation." More people were using methodologies developed during the early period. More electronic texts were being created and more projects using the same applications were started. Knowledge of what is possible had gradually spread through normal scholarly channels of communication, and more and more people had come across computers in their everyday life and had begun to think about what computers might do for their research and teaching.

The diffusion of knowledge was helped not only by *Computers and the Humanities* but also by a regular series of conferences. The 1970 symposium in Cambridge was the start of a biennial series of conferences in the UK, which became a major focal point for computing in the humanities. Meetings in Edinburgh (1972), Cardiff (1974), Oxford (1976), Birmingham (1978), and Cambridge (1980) all produced high-quality papers. The Association for Literary and Linguistic Computing was founded at a meeting in King's College London in 1973. Initially it produced its own *Bulletin* three times per year. It also began to organize an annual meeting with some invited presentations and by 1986 had a journal, *Literary and Linguistic Computing.* By the mid-1970s, another series of conferences began in North America, called the International Conference on Computing in the Humanities (ICCH), and were held in odd-numbered years to alternate with the British meetings. The British conference and the ALLC annual meetings gradually began to coalesce. They continued to concentrate on literary and linguistic computing with some emphasis on "linguistic", where they offered a forum for the

growing number of European researchers in what became known as corpus linguistics. ICCH attracted a broader range of papers, for example on the use of computers in teaching writing, and on music, art, and archaeology. The Association for Computers and the Humanities (ACH) grew out of this conference and was founded in 1978.

The requirements of humanities computing also began to be recognized within academic computing centers. Still in the days of mainframe computing, it was necessary to register to use any computing facilities and that registration provided an opportunity for academic computing staff to find out what users wanted and to consider providing some standard software that could be used by many different people. The second version of the COCOA concordance program in Britain was designed to be run on different mainframe computers for exactly this purpose (Berry-Rogghe and Crawford 1973). It was distributed to different computing centers in the mid-1970s and many of these centers designated one person to act as support. Dissatisfaction with its user interface coupled with the termination of support by the Atlas Laboratory, where it was written, led the British funding bodies to sponsor the development of a new program at Oxford University. Called the Oxford Concordance Program (OCP), this software was ready for distribution in 1982 and attracted interest around the world with users in many different countries (Hockey and Marriott 1979a, 1979b, 1979c, 1980). Other packaged or generic software also appeared at this time and significantly reduced the cost of a project in terms of programming support.

The need to avoid duplication of effort also led to consolidation in the area of text archiving and maintenance. With the advent of packaged software and the removal of the need for much programming, preparing the electronic text began to take up a large proportion of time in any project. The key driver behind the establishment of the Oxford Text Archive (OTA) in 1976 was the need simply to ensure that a text that a researcher had finished with was not lost. The OTA undertook to maintain electronic texts and, subject to the permission of the depositor and with appropriate copyright permissions, to make these texts available to anyone else who wanted to use them for academic purposes. It was the beginnings of a digital library, although nobody called it this initially, and its staff had to devise their own method of describing and documenting the material (Proud 1989). The amount of undocumented material highlighted the need for recognized procedures for describing electronic texts.

The OTA's approach was to offer a service for maintenance of anything that was deposited. It managed to do this for some considerable time on very little budget, but was not able to promote the creation of specific texts. Groups of scholars in some discipline areas made more concerted attempts to create an archive of texts to be used as a source for research. Notable among these was the *Thesaurus Linguae Graecae (TLG)* begun at the University of California Irvine and directed for many years by Theodore Brunner. Brunner raised millions of dollars to support the creation of a "databank" of Ancient Greek texts, covering all authors from Homer to about ad 600, some 70 million words (Brunner 1993). A complementary collection of Classical Latin was later produced by the Packard Humanities Institute, and together with the *TLG* gave scholars in classical studies a research resource that was unrivaled in other disciplines for many years. Only Old English scholars had access to a similar comprehensive, but smaller corpus with the completion of the Old English Corpus for the Dictionary of Old English (Healey 1989).

More centers for humanities computing were also established during this period. Some, for example the Norwegian Computing Center for the Humanities (now HIT) at Bergen, with substantial government support, incorporated a wide range of applications and projects. Others such as the Center for Computer Analysis of Texts (CCAT) at the University of Pennsylvania were more narrowly focused on the interests of the academics who had initially promoted them. Pockets of interest had become established around the world and scholars in those institutions on the whole enjoyed a good deal of support.

This period also saw the introduction of courses on various aspects of humanities computing. Some courses were given by staff within academic computing centers and concentrated mostly on the mechanics of using specific software programs. Others

looked more broadly at application areas. Those given by academics tended to concentrate on their own interests giving rise to student projects in the same application areas. A debate about whether or not students should learn computer programming was ongoing. Some felt that it replaced Latin as a "mental discipline" (Hockey 1986). Others thought that it was too difficult and took too much time away from the core work in the humanities. The string handling language SNOBOL was in vogue for some time as it was easier for humanities students than other computer languages, of which the major one was still Fortran.

There were some developments in processing tools, mostly through the shift from tape to disk storage. Files no longer had to be searched sequentially. For a time there were various technologies for organizing material in databases, some of which were very effective for humanities material (Burnard 1987b), but gradually the relational model prevailed. In mainframe implementations this presented a better structure within which historians and others working with material drawn from sources (rather than the sources themselves) could work. However, relational technologies still presented some problems for the representation of information that needed to be fitted into tables. At least two hardware devices were invented in the 1970s for assisting searching. One was implemented in David Packard's Ibycus computer, which was built to work with the *TLG* and some other classics material (Lancashire 1991: 204–5). The other was the Content Addressing File Store (CAFS), which worked on the British ICL computers (Burnard 1987a). The idea of transferring processing into the hardware was very attractive to humanities researchers who had to deal with large amounts of material, but it did not catch on in a big way, possibly because it was overtaken by advances in the speed of conventional hardware.

A glance through the various publications of this period shows a preponderance of papers based on vocabulary studies generated initially by concordance programs. The results were of interest either for some kinds of stylistic analyses or for linguistic applications. Increasingly complex mathematics were brought to bear on vocabulary counts, leaving some more humanities-oriented conference participants out in the cold. Apart from these, there was little really new or exciting in terms of methodology and there was perhaps less critical appraisal of methodologies than might be desirable. The important developments during this period lay more in support systems generated by the presence of more outlets for dissemination (conferences and journals) and the recognition of the need for standard software and for archiving and maintaining texts. Dissemination was concentrated in outlets for humanities computing and much less in mainstream humanities publications. It seems that we were still at a stage where academic respectability for computer-based work in the humanities was questionable and scholars preferred to publish in outlets where they were more likely to be accepted.

New Developments: Mid-1980s to Early 1990s

This period saw some significant developments in humanities computing. Some of these can be attributed to two new technologies, the personal computer and electronic mail. Others happened simply because of the increase of usage and the need to reduce duplication of effort.

At first there were several different and competing brands of personal computers. Some were developed for games, some were standalone word processors and could not be used for anything else, and others were specifically aimed at the educational market rather than for general use. Gradually IBM PCs and models based on the IBM architecture began to dominate, with Apple Macintoshes also attracting plenty of use, especially for graphics.

The personal computer is now a necessity of scholarly life, but in its early days it was considerably more expensive in relation to now and early purchasers were enthusiasts and those in the know about computing. The initial impact in humanities computing was that it was no longer necessary to register at the computer center in order to use a computer. Users of personal computers could do whatever they wanted and did not

necessarily benefit from expertise that already existed. This encouraged duplication of effort, but it also fostered innovation where users were not conditioned by what was already available.

By the end of the 1980s, there were three DOS-based text analysis programs: Word-Cruncher, TACT, and MicroOCP, all of which had very good functionality. Owners of personal computers would work with these at home and, in the case of WordCruncher and TACT, obtain instantaneous results from searches. MicroOCP was developed from the mainframe program using a batch concordance technique rather than interactive searching. However, the main application of personal computers was that shared with all other disciplines, namely word processing. This attracted many more users who knew very little about other applications and tended to assume that the functions within word processing programs might be all that computers could do for them.

The Apple Macintosh was attractive for humanities users for two reasons. Firstly, it had a graphical user interface long before Windows on PCs. This meant that it was much better at displaying non-standard characters. At last it was possible to see Old English characters, Greek, Cyrillic, and almost any other alphabet, on the screen and to manipulate text containing these characters easily. Secondly, the Macintosh also came with a program that made it possible to build some primitive hypertexts easily. HyperCard provided a model of file cards with ways of linking between them. It also incorporated a simple programming tool making it possible for the first time for humanities scholars to write computer programs easily. The benefits of hypertext for teaching were soon recognized and various examples soon appeared. A good example of these was the Beowulf Workstation created by Patrick Conner (Conner 1991). This presents a text to the user with links to a modern English version and linguistic and contextual annotations of various kinds. The first version of the Perseus Project was also delivered to the end user in HyperCard.

Networking, at least for electronic mail, was previously confined to groups of computer scientists and research institutes. By the mid-1980s, facilities for sending and receiving electronic mail across international boundaries were provided by most academic computing services. At the 1985 ALLC conference in Nice, electronic mail addresses were exchanged avidly and a new era of immediate communication began. Soon e-mail was being sent to groups of users and the ListServ software for electronic discussion lists was established. Ansaxnet, the oldest electronic discussion list for the humanities, was founded by Patrick Conner in 1986 (Conner 1992).

At the ICCH conference in Columbia, South Carolina, in spring 1987 a group of people mostly working in support roles in humanities computing got together and agreed that they needed to find a way of keeping in touch on a regular basis. Willard McCarty, who was then at the University of Toronto, agreed to look into how they might do this. On his return from the conference he discovered the existence of ListServ, and *Humanist* was born (McCarty 1992). The first message was sent out on May 7, 1987. McCarty launched himself into the role of editing what he prefers to call an "electronic seminar" and, except for a hiatus in the early 1990s when *Humanist* was edited from Brown University, has continued in this role ever since.

Humanist has become something of a model for electronic discussion lists. McCarty has maintained excellent standards of editing and the level of discussion is generally high. For those of us in Europe the regular early morning diet of three to six *Humanist* digests is a welcome start to the day. *Humanist* has become central to the maintenance and development of a community and it has made a significant contribution to the definition of humanities computing. Its archives going back to 1987 are a vast source of information on developments and concerns during this period and it was taken as an exemplar by the founders of the Linguist List, the key electronic forum for linguistics.

This period also saw the publication in print form of the only large-scale attempt to produce a bibliography of projects, software, and publications. Two volumes of the *Humanities Computing Yearbook (HCY)* were published. The first, edited by Ian Lancashire and Willard McCarty appeared in 1988 with some 400 pages. The second volume, for 1989–90, has almost 700 pages with a much better index. For several

years, until it began to get out of date, the *HCY* was an extremely valuable resource, fulfilling the role originally taken by the *Computers and the Humanities* Directory of Scholars Active, which had ceased to appear by the early 1970s. Preparing the *HCY* was a truly enormous undertaking and no further volumes appeared. By the early 1990s, the general consensus was that in future an online database would be a more effective resource. Although there have been various attempts to start something similar, nothing on a serious scale has emerged, and the picture of overall activity in terms of projects and publications is once again incomplete.

In terms of intellectual development, one activity stands out over all others during this period. In November 1987 Nancy Ide, assisted by colleagues in ACH, organized an invitational meeting at Vassar College, Poughkeepsie, to examine the possibility of creating a standard encoding scheme for humanities electronic texts (Burnard 1988). There had been various previous attempts to address the problem of many different and conflicting encoding schemes, a situation that was described as "chaos" by one of the participants at the Vassar meeting. Now, the time was ripe to proceed. Scholars were increasingly tired of wasting time reformatting texts to suit particular software and had become more frustrated with the inadequacies of existing schemes. In 1986, a new encoding method had appeared on the scene. The Standard Generalized Markup Language (SGML), published by ISO, offered a mechanism for defining a markup scheme that could handle many different types of text, could deal with metadata as well as data, and could represent complex scholarly interpretation as well as the basic structural features of documents.

Participants at the meeting agreed on a set of principles ("the Poughkeepsie Principles") as a basis for building a new encoding scheme and entrusted the management of the project to a Steering Committee with representatives from ACH, ALLC, and the Association for Computational Linguistics (Text Encoding Initiative 2001). Subsequently, this group raised over a million dollars in North America and oversaw the development of the Text Encoding Initiative (TEI) *Guidelines for Electronic Text Encoding and Interchange.* The work was initially organized into four areas, each served by a committee. Output from the committees was put together by two editors into a first draft version, which was distributed for public comment in 1990. A further cycle of work involved a number of work groups that looked at specific application areas in detail. The first full version of the TEI *Guidelines* was published in May 1994 and distributed in print form and electronically.

The size, scope, and influence of the TEI far exceeded what anyone at the Vassar meeting envisaged. It was the first systematic attempt to categorize and define all the features within humanities texts that might interest scholars. In all, some 400 encoding tags were specified in a structure that was easily extensible for new application areas. The specification of the tags within the *Guidelines* illustrates some of the issues involved, but many deeper intellectual challenges emerged as the work progressed. Work in the TEI led to an interest in markup theory and the representation of humanities knowledge as a topic in itself. The publication of the TEI *Guidelines* coincided with full-text digital library developments and it was natural for digital library projects, which had not previously come into contact with humanities computing, to base their work on the TEI rather than inventing a markup scheme from scratch.

Much of the TEI work was done by e-mail using private and public discussion lists, together with a fileserver where drafts of documents were posted. From the outset anyone who served on a TEI group was required to use e-mail regularly and the project became an interesting example of this method of working. However, participants soon realized that it is not easy to reach closure in an e-mail discussion and it was fortunate that funding was available for a regular series of face-to-face technical meetings to ensure that decisions were made and that the markup proposals from the different working groups were rationalized effectively.

Apart from major developments in personal computing, networking, and the TEI, the kind of humanities computing activities which were ongoing in the 1970s continued to develop, with more users and more projects. Gradually, certain application areas spun off from humanities computing and developed their own culture and dissemination

routes. "Computers and writing" was one topic that disappeared fairly rapidly. More important for humanities computing was the loss of some aspects of linguistic computing, particularly corpus linguistics, to conferences and meetings of its own. Computational linguistics had always developed independently of humanities computing and, despite the efforts of Don Walker on the TEI Steering Committee, continued to be a separate discipline. Walker and Antonio Zampolli of the Institute for Computational Linguistics in Pisa worked hard to bring the two communities of humanities computing and computational linguistics together but with perhaps only limited success. Just at the time when humanities computing scholars were beginning seriously to need the kinds of tools developed in computational linguistics (morphological analysis, syntactic analysis, and lexical databases), there was an expansion of work in computational and corpus linguistics to meet the needs of the defense and speech analysis community. In spite of a landmark paper on the convergence between computational linguistics and literary and linguistic computing given by Zampolli and his colleague Nicoletta Calzolari at the first joint ACH/ALLC conference in Toronto in June 1989 (Calzolari and Zampolli 1991), there was little communication between these communities, and humanities computing did not benefit as it could have done from computational linguistics techniques.

The Era of the Internet: Early 1990s to the Present

One development far outstripped the impact of any other during the 1990s. This was the arrival of the Internet, but more especially the World Wide Web. The first graphical browser, Mosaic, appeared on the scene in 1993. Now the use of the Internet is a vital part of any academic activity. A generation of students has grown up with it and naturally looks to it as the first source of any information.

Initially, some long-term humanities computing practitioners had problems in grasping the likely impact of the Web in much the same way as Microsoft did. Those involved with the TEI felt very much that HyperText Markup Language (HTML) was a weak markup system that perpetuated all the problems with word processors and appearance-based markup. The Web was viewed with curiosity but this tended to be rather from the outside. It was a means of finding some kinds of information but not really as a serious tool for humanities research. This presented an opportunity for those institutions and organizations that were contemplating getting into humanities computing for the first time. They saw that the Web was a superb means of publication, not only for the results of their scholarly work, but also for promoting their activities among a much larger community of users. A new group of users had emerged.

Anyone can be a publisher on the Web and within a rather short time the focus of a broader base of interest in humanities computing became the delivery of scholarly material over the Internet. The advantages of this are enormous from the producer's point of view. The format is no longer constrained by that of a printed book. Theoretically there is almost no limit on size, and hypertext links provide a useful way of dealing with annotations, etc. The publication can be built up incrementally as and when bits of it are ready for publication. It can be made available to its audience immediately and it can easily be amended and updated.

In the early to mid-1990s, many new projects were announced, some of which actually succeeded in raising money and getting started. Particularly in the area of electronic scholarly editions, there were several meetings and publications devoted to discussion about what an electronic edition might look like (Finneran 1996; Bornstein and Tinkle 1998). This was just at the time when editorial theorists were focusing on the text as a physical object, which they could represent by digital images. With the notable exception of work carried out by Peter Robinson (Robinson 1996, 1997, 1999) and possibly one or two others, few of these publications saw the light of day except as prototypes or small samples, and by the second half of the decade interest in this had waned somewhat. A good many imaginative ideas had been put forward, but once these reached the stage where theory had to be put into practice and projects were faced with the laborious work of entering and marking up text and developing software, attention began to turn elsewhere.

Debates were held on what to call these collections of electronic resources. The term "archive" was favored by many, notably the Blake Archive and other projects based in the Institute for Advanced Technology in the Humanities at the University of Virginia. "Archive" meant a collection of material where the user would normally have to choose a navigation route. "Edition" implies a good deal of scholarly added value, reflecting the views of one or more editors, which could be implemented by privileging specific navigation routes. SGML (Standard Generalized Markup Language), mostly in applications based on the TEI, was accepted as a way of providing the hooks on which navigation routes could be built, but significant challenges remained in designing and building an effective user interface. The emphasis was, however, very much on navigation rather than on the analysis tools and techniques that had formed the major application areas within humanities computing in the past. In the early days of the Web, the technology for delivery of SGML-encoded texts was clunky and in many ways presented a less satisfying user interface than what can be delivered with raw HTML. Nevertheless, because of the easy way of viewing them, the impact of many of these publishing projects was substantial. Many more people became familiar with the idea of technology in the humanities, but in a more limited sense of putting material onto the Web.

Although at first most of these publishing projects had been started by groups of academics, it was not long before libraries began to consider putting the content of their collections on the Internet. Several institutions in the United States set up electronic text or digital library collections for humanities primary source material, most usually using the OpenText SGML search engine (Price-Wilkin 1994). While this provides good and fast facilities for searching for words (strings), it really provides little more than a reference tool to look up words. Other projects used the DynaText SGML electronic book system for the delivery of their material. This offered a more structured search but with an interface that is not particularly intuitive.

A completely new idea for an electronic publication was developed by the Orlando Project, which is creating a History of British Women's Writing at the Universities of Alberta and Guelph. With substantial research funding, new material in the form of short biographies of authors, histories of their writing, and general world events was created as a set of SGML documents (Brown et al. 1997). It was then possible to consider extracting portions of these documents and reconstituting them into new material, for example to generate chronologies for specific periods or topics. This project introduced the idea of a completely new form of scholarly writing and one that is fundamentally different from anything that has been done in the past. It remains to be seen whether it will really be usable on a large scale.

The Internet also made it possible to carry out collaborative projects in a way that was never possible before. The simple ability for people in different places to contribute to the same document collections was a great advance on earlier methods of working. In the Orlando Project, researchers at both institutions add to a document archive developed as a web-based document management system, which makes use of some of the SGML markup for administrative purposes. Ideas have also been floated about collaborative editing of manuscript sources where people in different locations could add layers of annotation, for example for the Peirce Project (Neuman et al. 1992) and the Codex Leningradensis (Leningrad Codex Markup Project 2000). The technical aspects of this are fairly clear. Perhaps less clear is the management of the project, who controls or vets the annotations, and how it might all be maintained for the future.

The TEI's adoption as a model in digital library projects raised some interesting issues about the whole philosophy of the TEI, which had been designed mostly by scholars who wanted to be as flexible as possible. Any TEI tag can be redefined and tags can be added where appropriate. A rather different philosophy prevails in library and information science where standards are defined and then followed closely – this to ensure that readers can find books easily. It was a pity that there was not more input from library and information science at the time that the TEI was being created, but the TEI project was started long before the term "digital library" came into use. A few people made good contributions, but in the library community there was not the widespread range of many years' experience of working with electronic texts as in the

scholarly community. The TEI was, however, used as a model by the developers of the Encoded Archival Description (EAD), which has had a very wide impact as a standard for finding aids in archives and special collections.

An additional dimension was added to humanities electronic resources in the early 1990s, when it became possible to provide multimedia information in the form of images, audio, and video. In the early days of digital imaging there was much discussion about file formats, pixel depth, and other technical aspects of the imaging process and much less about what people can actually do with these images other than view them. There are of course many advantages in having access to images of source material over the Web, but humanities computing practitioners, having grown used to the flexibility offered by searchable text, again tended to regard imaging projects as not really their thing, unless, like the Beowulf Project (Kiernan 1991), the images could be manipulated and enhanced in some way. Interesting research has been carried out on linking images to text, down to the level of the word (Zweig 1998). When most of this can be done automatically we will be in a position to reconceptualize some aspects of manuscript studies. The potential of other forms of multimedia is now well recognized, but the use of this is only really feasible with high-speed access and the future may well lie in a gradual convergence with television.

The expansion of access to electronic resources fostered by the Web led to other areas of theoretical interest in humanities computing. Electronic resources became objects of study in themselves and were subjected to analysis by a new group of scholars, some of whom had little experience of the technical aspects of the resources. Hypertext in particular attracted a good many theorists. This helped to broaden the range of interest in, and discussion about, humanities computing but it also perhaps contributed to misapprehensions about what is actually involved in building and using such a resource. Problems with the two cultures emerged again, with one that was actually doing it and another that preferred talking about doing it.

The introduction of academic programs is another indication of the acceptance of a subject area by the larger academic community. For humanities computing this began to happen by the later 1990s although it is perhaps interesting to note that very few of these include the words "Humanities Computing" in the program title. King's College London offers a BA Minor in Applied Computing with a number of humanities disciplines, and its new MA, based in the Centre for Humanities Computing, is also called MA in Applied Computing. McMaster University in Canada offers a BA in Multimedia. The MA that the University of Virginia is soon to start is called Digital Humanities and is under the auspices of the Media Studies Program. The University of Alberta is, as far as I am aware, the first to start a program with Humanities Computing in its title, although the University of Glasgow has had an MPhil in History and Computing for many years.

As the Internet fostered the more widespread use of computers for humanities applications, other organizations began to get involved. This led to some further attempts to define the field or at least to define a research agenda for it. The then Getty Art History Information Program published what is in my view a very interesting Research Agenda for Networked Cultural Heritage in 1996 (Bearman 1996). It contains eight papers tackling specific areas that cover topics which really bridge across digital libraries, and humanities research and teaching. Each of these areas could form a research program in its own right, but the initiative was not taken further. Meanwhile the ALLC and ACH continued to organize a conference every year with a predominance of papers on markup and other technical issues. An attempt to produce a roadmap and new directions for humanities computing for the 2002 conference in Germany produced a useful survey (Robey 2002), but little new, and would perhaps have benefited from more input from a broader community. But how to involve other communities was becoming more of a problem in an era when many more electronic resources for the humanities were being developed outside the humanities computing community.

Conclusion

If one humanities computing activity is to be highlighted above all others, in my view it

must be the TEI. It represents the most significant intellectual advances that have been made in our area, and has influenced the markup community as a whole. The TEI attracted the attention of leading practitioners in the SGML community at the time when XML (Extensible Markup Language) was being developed and Michael Sperberg-McQueen, one of the TEI editors, was invited to be co-editor of the new XML markup standard. The work done on hyperlinking within the TEI formed the basis of the linking mechanisms within XML. In many ways the TEI was ahead of its time, as only with the rapid adoption of XML in the last two to three years has the need for descriptive markup been recognized by a wider community. Meanwhile, the community of markup theorists that has developed from the TEI continues to ask challenging questions on the representation of knowledge.

There are still other areas to be researched in depth. Humanities computing can contribute substantially to the growing interest in putting the cultural heritage on the Internet, not only for academic users, but also for lifelong learners and the general public. Tools and techniques developed in humanities computing will facilitate the study of this material and, as the Perseus Project is showing (Rydberg-Cox 2000), the incorporation of computational linguistics techniques can add a new dimension. Our tools and techniques can also assist research in facilitating the digitization and encoding processes, where we need to find ways of reducing the costs of data creation without loss of scholarly value or of functionality. Through the Internet, humanities computing is reaching a much wider audience, and students graduating from the new programs being offered will be in a position to work not only in academia, but also in electronic publishing, educational technologies, and multimedia development. Throughout its history, humanities computing has shown a healthy appetite for imagination and innovation while continuing to maintain high scholarly standards. Now that the Internet is such a dominant feature of everyday life, the opportunity exists for humanities computing to reach out much further than has hitherto been possible.

References for Further Reading

Bearman, D., (ed.) (1996). *Research Agenda for Networked Cultural Heritage*. Santa Monica, CA: Getty Art History Information Program.

Berry-Rogghe, G. L. M. and T. D. Crawford (1973). *Developing a Machine-independent Concordance Program for a Variety of Languages*. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (eds.), *The Computer and Literary Studies* (pp. 309–16). Edinburgh: Edinburgh University Press.

Bessinger, J. B. and S. M. Parrish (1965). *Literary Data Processing Conference Proceedings*. White Plains, NY: IBM.

Bornstein, G. and T. Tinkle (1998). *The Iconic Page in Manuscript, Print, and Digital Culture*. Ann Arbor: University of Michigan Press.

Brown, S., S. Fisher, P. Clements, K. Binhammer, T. Butler, K. Carter, I. Grundy, and S. Hockey (1997). *SGML and the Orlando Project: Descriptive Markup for an Electronic History of Women's Writing*. *Computers and the Humanities* 31: 271–84.

Brunner, T. F. (1993). *Classics and the Computer: The History of a Relationship*. In J. Solomon (ed.), *Accessing Antiquity: The Computerization of Classical Studies* (pp. 10–33). Tucson: University of Arizona Press.

Burnard, L. (1987a). *CAFS: A New Solution to an Old Problem*. *Literary and Linguistic Computing* 2: 7–12.

Burnard, L. (1987b). *Principles of Database Design*. In S. Rahtz (ed.), *Information Technology in the Humanities* (pp. 54–68). Chichester: Ellis Horwood.

Burnard, L. (1988). *Report of Workshop on Text Encoding Guidelines*. *Literary and Linguistic Computing* 3: 131–3.

Burton, D. M. (1981a). *Automated Concordances and Word Indexes: The Fifties*. *Computers and the Humanities* 15: 1–14.

Burton, D. M. (1981b). *Automated Concordances and Word Indexes: The Early Sixties and the Early Centers*. *Computers and the Humanities* 15: 83–100.

Burton, D. M. (1981c). *Automated Concordances and Word Indexes: The Process, the Programs, and the Products*. *Computers and the Humanities* 15: 139–54.

Burton, D. M. (1982). *Automated Concordances and Word Indexes: Machine Decisions and Editorial Revisions*. *Computers and the Humanities* 16: 195–218.

Busa, R. (1974-). *Index Thomisticus*. Stuttgart: Frommann-Holzboog.

Busa, R. (1980). *The Annals of Humanities Computing: The Index Thomisticus*. *Computers and the Humanities* 14: 83–90.

Busa, R., (ed.) (1992). *Thomae Aquinatis Opera Omnia Cum Hypertextibus in CD-ROM*. Milano: Editoria Elettronica Editel.

Busa, R. (1999). *Picture a Man.… Busa Award Lecture, Debrecen, Hungary, July 6, 1998*. *Literary and Linguistic Computing* 14: 5–9.

Calzolari, N. and A. Zampolli (1991). *Lexical Databases and Textual Corpora: A Trend of Convergence between Computational Linguistics and Literary and Linguistic Computing*. In S. Hockey, N. Ide, and I. Lancashire (eds.), *Research in Humanities Computing 1*: Selected Papers from the ALLC/ACH Conference, Toronto, June 1989 (pp. 272–307). Oxford: Clarendon Press.

Conner, P. W (1991). *The Beowulf Workstation: One Model of Computer-assisted Literary Pedagogy*. *Literary and Linguistic Computing* 6: 50–8.

Conner, P. W (1992). *Networking in the Humanities: Lessons from Ansaxnet*. *Computers and the Humanities* 26: 195–204.

De Tollenaere, F. (1973). *The Problem of the Context in Computer-aided Lexicography*. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (eds.), *The Computer and Literary Studies* (pp. 25–35). Edinburgh: Edinburgh University Press.

Ellegård, A. (1962). *A Statistical Method for Determining Authorship: The Junius Letters 1769–1772*. Gothenburg: Gothenburg Studies in English.

Finneran, R. J. (1996). *The Literary Text in the Digital Age*. Ann Arbor: University of Michigan Press.

Gorcy, G. (1983). *L'informatique et la mise en oeuvre du trésor de la langue française (TLF), dictionnaire de la langue du 19^{e} et du 20^{e} siècle (1789–1960)*. In A. Cappelli and A. Zampolli (eds.), *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries: Proceedings of the European Science Foundation Workshop, Pisa* 1981. *Linguistica Computazionale* III (pp. 119–44). Pisa: Giardini.

Hamilton-Smith, N. (1971). *A Versatile Concordance Program for a Textual Archive*. In R. A. Wisbey (ed.), *The Computer in Literary and Linguistic Research* (pp. 235–44). Cambridge: Cambridge University Press.

Healey, A. (1989). *The Corpus of the Dictionary of Old English: Its Delimitation, Compilation and Application*. Paper presented at the Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary. Oxford, September, 1989.

Hockey, S. (1986). *Workshop on Teaching Computers and the Humanities Courses*. *Literary and Linguistic Computing* 1: 228–9.

Hockey, S. and I. Marriott (1979a). *The Oxford Concordance Project (OCP) – Part 1*. *ALLC Bulletin* 7: 35–43.

Hockey, S. and I. Marriott (1979b). *The Oxford Concordance Project (OCP) – Part 2*. *ALLC Bulletin* 7: 155–64.

Hockey, S. and I. Marriott (1979c). *The Oxford Concordance Project (OCP) – Part 3*. *ALLC Bulletin* 7: 268–75.

Hockey, S. and I. Marriott (1980). *The Oxford Concordance Project (OCP) – Part 4*. *ALLC Bulletin* 8: 28–35.

Holmes, D. I. and R. S. Forsyth (1995). *The Federalist Revisited: New Directions in Authorship Attribution*. *Literary and Linguistic Computing* 10: 111–27.

Kiernan, K. S. (1991). *Digital Image Processing and the Beowulf Manuscript*. *Literary and Linguistic Computing* 6: 20–7.

Lancashire, I., (ed.) (1991). *The Humanities Computing Yearbook 1989–90: A Comprehensive Guide to Software and Other Resources*. Oxford: Clarendon Press.

Lancashire, I. and W. McCarty, (eds.) (1988). *The Humanities Computing Yearbook 1988*. Oxford: Clarendon Press.

Leningrad Codex Markup Project (2000). *Project "EL": The XML Leningrad Codex*. Available at: <http://www.leningradensis.org>, accessed May 15, 2003.

Lord, R. D. (1958). *Studies in the History of Probability and Statistics: viii. de Morgan and the Statistical Study of Literary Style*. *Biometrika* 45: 282.

McCarty, W. (1992). *Humanist: Lessons from a Global Electronic Seminar*. *Computers and the Humanities* 26: 205–22.

Mendenhall, T. C. (1901). *A Mechanical Solution of a Literary Problem*. *The Popular Science Monthly* 60: 97–105.

Morton, A. Q. (1965). *The Authorship of the Pauline Epistles: A Scientific Solution*. Saskatoon: University of Saskatchewan.

Morton, A. Q. and Winspear, A. D. (1971). *It's Greek to the Computer*. Montreal: Harvest House.

Mosteller, F. and D. L. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

Neuman, M., M. Keeler, C. Kloesel, J. Ransdell, and A. Renear (1992). *The Pilot Project of the Electronic Peirce Consortium (abstract)*. *ALLC-ACH92 Conference Abstracts and Program* (pp. 25–7). Oxford.

Parrish, S. M. (1962). *Problems in the Making of Computer Concordances*. *Studies in Bibliography* 15: 1–14.

Price-Wilkin, J. (1994). *Using the World Wide Web to Deliver Complex Electronic Documents: Implications for Libraries*. *The Public-Access Computer Systems Review* 5: 5–21. <http://jpw.umdl.umich.edu/pubs/yale.html>, accessed July 21, 2004.

Proud, J. K. (1989). *The Oxford Text Archive*. London: British Library Research and Development Report.

Robey, D. (2002). *New Directions in Humanities Computing*, <http://www.uni-tuebingen.de/zdv/zrkinfo/pics/aca4.htm>, accessed May 15, 2003.

Robinson, P., (ed.) (1996). *Geoffrey Chaucer: The Wife of Bath's Prologue on CD-ROM*. Cambridge: Cambridge University Press.

Robinson, P. M. W. (1997). *New Directions in Critical Editing*. In K. Sutherland (ed.), *Electronic Text: Investigations in Method and Theory* (pp. 145–71). Oxford: Clarendon Press.

Robinson, P. M. W. (1999). *New Methods of Editing, Exploring and Reading The Canterbury Tales*. http://www.cta.dmu.ac.uk/projects/ctp/desc2.html, accessed May 14, 2003.

Russell, D. B. (1967). *COCOA - A Word Count and Concordance Generator for Atlas*. Chilton: Atlas Computer Laboratory.

Rydberg-Cox, J. A. (2000). *Co-occurrence Patterns and Lexical Acquisition in Ancient Greek Texts*. *Literary and Linguistic Computing* 15: 121–30.

Text Encoding Initiative (2001). *Text Encoding Initiative*, http://www.tei-c.org, accessed May 15, 2003.

Tweedie, F. J., S. Singh, and D. I. Holmes (1996). *Neural Network Applications in Stylometry: The Federalist Papers*. *Computers and the Humanities* 30: 1–10.

Wisbey, R. (1963). *The Analysis of Middle High German Texts by Computer: Some Lexicographical Aspects*. *Transactions of the Philological Society*, 28–48.

Wisbey, R. A., (ed.) (1971). *The Computer in Literary and Linguistic Research*. Cambridge: Cambridge University Press.

Zweig, R. W. (1998). *Lessons from the Palestine Post Project*. *Literary and Linguistic Computing* 13: 89–97.