The role of humanities computing: experiences and challenges

Short, Harold

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

The Role of Humanities Computing : Experiences and Challenges

Harold Short *

Abstract: Dued to the celebration of the thirtieth anniversary of the *Department for Literary and Documentary Data Processing* in Tuebingen this article is written. It gives an overview of humanities computing developments since the formation of this Research-Department. The paper is devided into three parts. First, the experiences in humanities computing are reviewed. For these purposes the author points out various aspects of the development and exploitation of scholarly materials using computers, considering some of the current work to create new tools for research. This chapter is followed by the discussion of some of the key challenges of this century, by that humanities computing and the scholarship, of which it is a part, are faced with. Finally, the author gives a summary of what in his opinion would be the key roles of humanities computing in the future.

Introduction

We are meeting today to acknowledge and celebrate the more than 30 years of humanities computing here in Tuebingen, and the thirtieth anniversary of the Department for Literary and Documentary Data Processing (Abteilung Literarische und Dokumentarische Datenverarbeitung): http://www.uni-tuebingen.de/zdv/tustep/.

---

* Aus dem Protokoll des *80. Kolloquiums* über die Anwendung der Elektronischen Datenverarbeitung in den Geisteswissenschaften an der Universität Tübingen vom 18. November 2000 (Revised version: May 2002).

Address all communications to Harold Short, King's College London, Centre for Computing in the Humanities, Strand, London WC2R 2LS, United Kingdom.
E-mail: Harold.Short@kcl.ac.uk.

It is a great privilege to be here to join your celebration. I had planned to be here in any case, but to be sitting out there with you rather than standing here. So although I was honoured by Professor Ott's invitation to take Professor Zampolli's place, I also found it extremely daunting, all the more so when I realised that the person who had spoken at the *20th* anniversary was that great pioneer of our field, Roberto Busa SJ.[1]

It is much to be regretted that our esteemed colleague Professor Zampolli could not be present, and I echo Professor Ott's good wishes to him. He would have been a most appropriate speaker, for he attended that famous colloquium here in 1960[2] with Father Busa, and has been closely involved in the whole range of humanities computing developments over the intervening years. He was a student of Busa and a collaborator with Ott and Wisbey and others in the founding of the Association for Literary and Linguistic Computing (ALLC: http://www.allc.org) in 1973. I cannot begin to have either the depth or the breadth of his understanding across the field of humanities computing.

This is, I believe, an exciting time for anyone involved in the application of computing in the humanities disciplines. Information and communications technologies are evolving at a pace almost impossible to keep up with and difficult even to comprehend. There are new areas of research and activity that impinge on us more and more in fields near and far: information science, computer science, engineering, cognitive science, not to mention the wider cultural heritage sector, with which humanities scholarship has had more traditional ties. Somehow we have to make ourselves aware of and make sense of all of this.

Experiences old and new

Our experiences in humanities computing now go back more than 50 years, each decade more packed with change and innovation than the one before. In what must necessarily be a brief and highly selective overview, I will concentrate on the scholarly creation and use of 'digital resources' (to use the current term), and will refer to a number of projects and institutions involved in this work. This should certainly NOT be taken to indicate a lower regard for other kinds of scholarly use of computers; far from it.

The focus of my discussion is what I am calling 'hybrid resources for scholarship'. The creation of such 'resources' - the product of scholarship and created

[1] Kolloquium 50 on 24 November 1990. Busa's paper was entitled 'Half a Century of Literary Computing: Towards a "New" Philology'. See *Literary and Linguistic Computing* 7 (1992) 1, 68-73.

[2] This Kolloquium, entitled "Internationales Kolloquium über maschinelle Methoden der Literarischen Analyse und der Lexikographie", was held in the University of Tuebingen on 24 November 1960.

for the use of scholars - has been a key activity throughout our fifty years. Roberto Busa set out on his computing path with the purpose of producing a concordance and index to the works of Thomas Aquinas.[3] It was a *work* of substantial research, and a *product* of research that itself has become an invaluable *tool* of research.

The record of scholarly publication here in the ZDV/ALDDV in Tuebingen is remarkable, and it could be characterised in a similar way. The record of achievement and of far-sighted technical design is one of which he and his colleagues in LDDV, and the University of Tuebingen more generally, should be proud. It is an achievement that is known and admired far beyond this institution and this city.

Professor Ott has told me that more than 1800 volumes have been prepared with TUSTEP. The *editions* page on the TUSTEP web site http://www.uni-tuebingen.de/zdv/tustep/ tustep_eng.html needs an index for its several hundreds of entries, and that is before you get on to the lists of bibliographies, indexes, concordances, dictionaries and other reference works in whose production it has been used.

The changes over the years in the character of the publications, in the TUSTEP software itself, and no less in the papers given in this Kolloquium, reflect developing technologies and changing expectations, and the increasing intermixture of paper and electronic forms. At first a key role of the technology was to make possible or enhance the paper publication. As the technologies have developed, and the scholarly horizons have changed to take account of them, so the publication forms have developed. In many cases now it is the digital form that is the key resource for scholars, and if there are paper publications at all, they are supplementary to the electronic one.

Hence my use, above, of the term 'hybrid' - encompassing research resources in both digital and non-digital forms, not just the purely digital. In using the term I seek to include not only a mixed form of publication, but also the paper or other material form of the sources on which the digital forms may be based. Thus I do not see 'hybrid' as merely an intermediate stage between 'paper' and 'digital'; rather as something long term, perhaps even in some way 'ideal', in which a paper or material original is preserved and treasured for what it is, and the electronic is exploited for what it makes possible.

One important characteristic of the work of Busa, Ott and many others has been that it falls firmly within the existing traditions of textual scholarship. What has been developed is built on solid foundations and is highly practical, although this does not mean it has failed to be innovative. It has made possible many things that were not feasible before. Busa's *Index Thomisticus* could be

[3] Commonly known as the *Index Thomisticus: The Busa Edition (1974-1980): S. Thomae Aquinatis Opera Omnia; ut sunt in indice thomistico, additis 61 scriptis ex aliis medii aevi auctoribus*. Fromann-Holzboog, Stuttgart-Bad Canstatt, 1980; available on CD-ROM as a supplement to the 56-volume printed edition.

completed in a matter of decades, rather than of centuries or not at all; concordances have become matters of routine rather than a life's work, a point made by Michael Sperberg-McQueen in his paper in this Kolloquium five years ago.[4] TUSTEP has brought new possibilities of rigour, consistency and comprehensiveness to the preparation and production of scholarly editions, and has made possible new kinds of editions.

Textual studies: thinking with mark-up

Of course not everyone uses TUSTEP. Since all processing of texts by computer involves some form - perhaps several forms - of mark-up (or encoding), it was natural that as the volume of work of this kind increased, so the question of standards should arise. This in turn gave birth to the Text Encoding Initiative (TEI), one of the most remarkable projects of its kind to have been undertaken. My authority for this is not a computing humanist nor anyone directly involved, but Jon Bosak, Chair of the XML Work Group of the World Wide Web Consortium.[5] The intellectual basis of TEI as a document grammar system was a key part of the Sperberg-McQueen paper here 5 years ago, so I will touch instead on some key practical consequences of its development and use.
The TEI is remarkable for at least three reasons.

- First, it it does indeed provide a basis for ensuring that texts can be transferred between different hardware and software platforms without loss of data, not only providing a measure of 'future-proofing' against hardware and software changes, but also enabling scholars to exchange and share their encoded texts.
- Second, it engaged some of the finest scholarly minds in a common endeavour over more than a decade, a process which yielded particular insights into the ways in which mark-up may be of much greater value than 'mere' longevity in a variety of scholarly endeavours.
- Third, the work of the project led directly to its North American editor, Michael Sperberg-McQueen, taking a leading role in that W3C Work Group that defined the XML (eXtensible Mark-up Language) standard.

The TEI is continuing, as a Consortium co-hosted at Virginia, Brown, Bergen and Oxford (http://www.tei-c.org/), and Sperberg-McQueen now works for W3C. As the wider XML 'revolution' gathers pace, there are signs that some of the long-term significance of the TEI will be related to XML, and the opportunities it is starting to bring to textual scholarship, not only in burgeoning quan-

[4] Kolloquium 65: 18 November 1995. See *Literary and Linguistic Computing* 12 (1997) 1, 53-60.

[5] Bosak characterised the TEI in this way during his closing Keynote talk at the 'TEI 10' tenth aniversary conference, held at Brown University, 14-16 November 1997. http://www.stg.brown.edu/conferences/tei10

tities of encoded texts (of which more later), but also in the development of new tools to exploit them.

As you are no doubt aware, in order to produce a basic TEI-conformant text it is necessary to carry out only what is sometimes termed 'shallow' or 'light' mark-up - enough essentially to capture the structure of the text - the chapters and paragraphs of a novel, the acts, scenes and lines of a play, the stanzas and lines of a poem. TEI encoding can be used, of course, to do much more, but that is the minimum. Let us look next at an example of very 'deep' mark-up, which, although it does not actually use TEI (it began before the TEI standards were defined), illustrates some of the intellectual challenges and potential benefits of deep encoding in textual scholarship.

For the past eleven years my friend and colleague Willard McCarty has been preparing an *Analytical Onomasticon to the Metamorphoses of Ovid*: http://ilex.cc.kcl.ac.uk/analyticalonomasticon/index.htm
There is not time here to attempt any detailed description of the work, but to quote McCarty, it is

> "in essence a systematic, disciplined means of discovering and following paths of association through the *Metamorphoses* - specifically those formed by shared names and what I call 'onomastic profiles' ". - (Unpublished. Quoted with permission.)

It is a 'book of names' of an entirely new kind, in which a systematic attempt is made to capture all references to 'persons', whether directly by proper name or by other appellative devices, direct or indirect. Given the subject matter of the poem, the poetic process of personification is of particular interest.

Classical scholars will judge how effective the *Onomasticon* is as a research tool when it is published. However, I want to focus on its methodology. What McCarty set out to do required the construction of a metalanguage - a complex set of tags - that would enable him to mark up in the text all the words and phrases that he judges to have an appellative function. The marked up text constituted a 'model' of the poem and its appellative devices; as the work progressed, there was a process of change, both in the tag set itself and in the tagging, thereby changing the 'model'. Thus when someone uses the resource in its published electronic form, what they will see will reflect McCarty's interpretation of Ovid's poetics. Reflections of this kind are not themselves new, of course - all publications represent the interpretations of their authors or editors. However, what *will* be new is that the basis of his interpretation, his model, will be entirely and comprehensively explicit. His critics will be able to take issue with anything from the encoding of a single word to his entire metalanguage scheme. What is more, it will be possible for other scholars not only to 'replicate' his results, but also to change the mark-up and thence to produce new

results - they will change the model, and so will be able to generate new interpretations.
There are two other aspects of the work that are worthy of comment, both raised by McCarty himself at various times.

- The first is to do with his experience of the encoding process. The machine has no intelligence, and is therefore merciless in showing up inconsistencies and contradictions in the mark-up scheme. Thus the process itself becomes a sigificant tool for thinking about the poem and its poetic devices. It also is a very good illustration of the modelling-failure cycle, which was referred to or even emphasized, I noticed, in the papers given in this Kolloquium by Busa (Kolloquium 50: 24 November 1990), Sperberg-McQueen (Kolloquium 65: 18 November 1995) and Susan Hockey (Kolloquium 74: 5 December 1998). Each 'failure' gave McCarty new insights both into his scheme and into Ovid's work, and provided a basis for refining his model.
- The second is to do with the commitment of time and energy required by such an undertaking. Many - perhaps most - words in the poem have more than one tag attached to them; in such a long poem, this represents substantial labour. Deep encoding is not for everyone! However it illustrates in a very practical way one of the fundamental potentialities of humanities computing - and I am not aware of many undertakings of its specific kind. I may of course be wrong, but I believe that in time it will come to be seen as a piece of seminal work, perhaps to rank even with Busa's *Index Thomisticus* - but perhaps at this point I should remind you also that McCarty is a close friend and colleague!

These remarks should not be understood to imply that scholars use the computer as a tool for thinking *only* through deep encoding. From the beginning the best and most intelligent use of computing has enabled new ways of thinking about the materials of scholarship, and the iterative process of modelling and failure characterises much if not all scholarly work with computers.[6]

Databases and structured data

Another area of activity in which a similar phenomenon may be observed is historical studies. I will not attempt to encompass all of such a broad field. I noticed that one of your speakers here was Professor Manfred Thaller, who spoke to you about the CLIO system developed by him and his colleagues (Kolloquium 38: 22 November 1986). He will have given you a much wider overview of historical computing than I could possibly do. Instead I will focus

[6] See also Unsworth, J.: "The Importance of Failure," in *The Journal of Electronic Publishing*, 3.2 (December, 1997).

on one particular kind of research, that of prosopography, in which the Centre for Computing in the Humanities (http://www.kcl.ac.uk/cch) at King's College has some direct involvement. We are collaborators in three major projects of this kind, funded by the UK's Arts and Humanities Research Board: these are three projects:

- the *Prosopography of the Byzantine Empire*, (http://www.kcl.ac.uk/cch /PBE)
- the *Prosopography of Anglo-Saxon England*, (http://www.kcl.ac.uk/cch /pase)
- the *Clergy of the Church of England Database* (http://www.kcl.ac.uk /cch/cce)

A fourth project, the *Prosopography of Roman Egypt*, is on temporary hold. In each case information is recorded within a formalised structure - in these projects the data is managed by relational database software - which reflects the scholars' view of what is important in the subject matter; or, to be more precise, what is both important and also amenable to being forced into a structure of this kind, which is an important qualification. Where this approach is appropriate, the database tools make possible the rapid retrieval of data in many combinations, the asking of more or less complex questions, and the selection of sets of data that may reveal patterns or discontinuities, and that demonstrate or suggest links between different data elements. As with all the best scholarship, it may be at its most useful when it raises new questions, more so perhaps than when it helps to provide 'answers'.

Projects such as these pose technical as well as scholarly challenges. The technical approach adopted in each involves the development of a 'data collection database' to make the gathering of information as efficient as possible with respect to the current scope of an individual researcher, or to the characteristics of a particular source type; the development of a 'master database' into which the collected data is uploaded and integrated; and procedures for producing richly varied means of accessing and manipulating the data.

There is another database project at King's College which has a very different scholarly purpose - a literary one. Dr David Yeandle is creating a line-by-line bibliography for Wolfram von Eschenbach's *Parzival*.[7] Database entries are created for all published work on the poem, recording the line numbers and thematic aspects addressed. This makes it possible to access the bibliographic information by line number, theme, author and year.

Although this project uses database technology rather than mark-up, there are parallel's with McCarty's *Onomasticon*: in the comprehensiveness of its aims,

[7] Yeandle, David N.: *Stellenbibliographie zum "Parzival" Wolframs von Eschenbach für die Jahrgänge 1984-1996* Niemeyer Verlag, Tübingen, to be published on CD-ROM, 2002. Details of the project may be found at: http://www.kcl.ac.uk/kis/schools/hums/german /parzive.html

and the level of detail at which the work is done. In both cases, also, 'publication' makes sense only in electronic form - although certain kinds of paper output might be useful as 'spin-off' publications. In this they are examples of 'second generation digital resources', to which I will return later.

Computational and Corpus Linguistics

Philology has traditionally encompassed both literary and linguistic scholarship, and it is not surprising that the origins of the Association for Literary and Linguistic Computing in the early 1970s lie in a time of awareness of parallels and overlaps in the application of computing in linguistic and literary studies. Some of the developments in computational linguistics during the 80s and 90s were based in and driven by computer science departments, with an emphasis on theoretical linguistics, and these seemed remote from the interests of literary scholars, so 'literary computing' and 'linguistic computing' appeared to diverge to some extent. More recently, however, the emergence of 'corpus linguistics' and of corpus-based research more generally, as well as continuing work in applied linguistics, have fostered a welcome process of re-convergence. In part this is based on a recognition that theoretical and practical issues in corpus building and corpus use are common to linguistic and literary studies, and in part on an understanding of how the tools developed for linguistic applications may have a significant role in literary research. Professor Zampolli's institute in Pisa, the *Istituto di Linguistica Computazionale*, is just one example, among many, of where tools of this kind have been developed, and where the 'linguistic' and the 'literary' have not been divorced (http://www.ilc.pi.cnr.it/).

There is also a wider social context for this process of re-convergence. Corpus linguistics provides the foundation of a great deal of work in the field of 'human language technologies', which today has major cultural, political and commercial significance, notably in the European context, but also far beyond, on a truly international scale. The European Language Resource Association (ELRA: http://www.icp.inpg.fr/ELRA/), in whose establishment Antonio Zampolli has played a major role, demonstrates the vibrancy and potential of current work in this area, for example in the range of resources and tools created under its aegis. Professor Zampolli could tell you much more about these developments than I can, but they remain as central to the experiences of 'humanities computing' as they were at the start of our journey, in the work of Father Busa.

Mixed media resources

Mixed media resources have become increasingly important to humanities scholars as the technologies that enable them have developed, including increasingly sophisticated techniques to manipulate images, the decreasing cost

of storage that make it possible for very large files to be manipulated even on personal computers, and the increasing bandwidth of the networks that make possible the rapid transfer of high data volumes.

As my first example I would draw your attention to the Blake Archive (http://www.iath.virginia.edu/blake/), based in the Institute for Advanced Technology in the Humanities (IATH: http://www.iath.virginia.edu) at the University of Virginia. It is a representation of the work of someone who was both a poet and an artist, and whose work can be understood best when the poetry and the art are seen and studied in an integrated way. Thus the intellectual requirement finds a very appropriate match in the technical possibilities.

One cannot help being struck by the degree and quality of intellectual input required in the creation of a resource of this kind if it is to be of value to a scholarly audience, even if on a superficial level its ease of use makes it accessible to a much wider non-specialist audience. For example, substantial 'metadata' has been created to describe the image content, which makes possible content-based searches of the images and also thematic cross-referencing between images and poems.

Other aspects of relevance to our present discussion include the 'modelling' of the data that underlies the resource, the range of sources and contributors it draws on, and the collaborative basis of its development. I shall return to some of these matters later.

As my next example, I have chosen a project at the Courtauld Institute of Art, the *Corpus of Romanesque Sculpture in Britain and Ireland* (CRSBI: http://www.crsbi.ac.uk).

The resource consists of photographs of romanesque sculpture and with expert descriptions and commentaries. The researchers are volunteer art historians, and the work of the project is overseen by a committee of eminent scholars. The researchers take the photographs and prepare the written material; the photographs and texts are then sent in to the project office, where they are edited and prepared for on-line publication by the Editor and the Research Officer. As the material is received and processed, the details are recorded in a central database. The images are scanned - at a high resolution for preservation purposes, with lower resolution (JPEG) versions produced for web display purposes (thumbnail and full-screen). The texts are edited, with XML tags inserted to reflect the carefully designed structure of information the scholars are asked to produce, as well as to identify specialised terms so they can be linked later to a glossary. The database, image and XML data are then used to display integrated HTML materials on the web site, by means of a set of (Perl) scripts. I sketch the process because I believe it is typical of many multi-media digital resource projects currently in progress.

Another such project at the Courtauld, the *Corpus Vitrearum Medii Aevi* (CVMA: http://www.cvma.ac.uk), is digitising photographs of medieval stained glass and adding a great deal of metadata in order to create another

searchable on-line database that will be valuable to scholars as well as having wider appeal. The CVMA is in fact an international project, with well-established paper publication series of scholarly commentaries, and the pilot project is enabling the CVMA to explore how such commentaries could in the future be integrated with the image archive.

Among the things I find interesting about these three projects - and many others like them - is the similarity in the intellectual and technical challenges posed by taking different kinds of source material and shaping them in an intellectually rigorous way into a whole that aims to provide a basis for new scholarship. But they have other important characteristics. I selected the Blake Archive because it provides a practical interdisciplinary framework for literary scholars and art historians; CRSBI because it enables important art historical evidence to preserved that would otherwise be lost; CVMA because it has the potential to bring together into a single unified resource images of all the UK's medieval stained glass (Europe's too if the international project adopts the same approach across its member countries).

These characteristics - of bringing source materials together in 'virtual' collections, of bridging the gaps between disciplines, and of providing access to remote or fragile resources - are in one or more respects true of a great many of the scholarly multi-media projects now under way, and in the latter respect at least of the hundreds of digitisation projects being undertaken by museums, galleries and libraries all round the world. All such resources are likely to be used by scholars for research and teaching, as part of the global digital library. The management of this 'library' to ensure access and long-term availability is far from straightforward, and this is the first of the *challenges* to which we'll turn our attention in a moment.

Bringing to a close this rapid and selective survey of the 'experiences' of humanities computing, I should emphasise the fact that many major areas of activity have not been touched on, such as the study of literary style[8] and authorship attribution. Let me repeat that their omission is only because I chose to take a 'digital resources' perspective for this talk, not because I regard them as less important for humanities computing or for scholarship.

[8] cf John Burrows and his seminal work on Jane Austen, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford, Clarendon Press, 1987, and the substantial body of work by Burrows, McKenna, Craig and others at the Centre for Literary and Linguistic Computing at the University of Newcastle, New South Wales. For a recent overview of work in this area, see Holmes, D.I.: *The evolution of stylometry in humanities scholarship* in: *Literary and Linguistic Computing* vol 13 no 3 (1998), pp. 111-117.

Challenges

Many of the challenges that face us in the new century arise from the rapidity of technological change and the loss of stability in institutions and practices, and in society at large, that follow from this. Among the most complex of these is the management and preservation of the ever-growing range of digital resources. From a scholarly point of view what is particularly important is the integration of the digital and the non-digital in what I described earlier as the 'hybrid' library.

The hybrid library

All of the resources we have so far looked at might be termed 'second generation' digital resources, according to criteria proposed by John Unsworth in the opening keynote address at the Digital Resources for the Humanities Conference (DRH) 2000 in Sheffield.[9] In his paper Unsworth suggested that the distinguishing features of 'second-generation resources' are likely to include the following: collaborative; multi-disciplinary; multi-media; multi-technology; large-scale; complex; long-term (both in terms of their creation and their intended usefulness).[10] This formulation is useful, although clearly not exclusive, and in his paper Unsworth identified a number of important issues that arise in the conception, planning, project management, development, dissemination, management, and preservation of these resources. I should add that it would be against the inclusive spirit of Unsworth's paper to see his proposal as a yardstick for deciding whether some piece of work is 'old-fashioned' or 'modern'; rather its importance lies in its systematic analysis of the issues raised by new technical possibilities, and in his proposed agenda of research that follows from the analysis. His focus is on:

- the scholarly use of primary resources in digital form
- the adoption by libraries of what he calls 'born-digital' resources
- the new interactions needed by scholars, publishers and libraries in the co-creation and dissemination of scholarly digital resources

In part, Unsworth's proposals arise from a recognition that the scale of the operation is changing, that we are approaching a threshold that necessitates new thinking. The Andrew W. Mellon Foundation is funding Unsworth's group to carry out research in this area: *Supporting Digital Scholarship* (http://www.iath.virginia.edu/sds/).

---

[9] *Digital Resources for the Humanities* (DRH) is an annual conference that specifically addresses the scholarly, information and technical issues related to the creation, management, use and preservation of digital resources in the humanities: http://www.drh.org.uk. DRH2000 was held at the University of Sheffield in September 2000.

[10] This is my own summary and interpretation of what he said, rather than a quotation from a published version of the paper, which has not appeared as yet.

It may also be worth mentioning: the UK's Arts and Humanities Data Service,[11] which supports the creation of scholarly digital resources, in part by recommending technical standards and specifying metadata structures to ensure wide access and long-term preservation; and the FEDORA architecture for a 'digital objects repository', developed by computer scientists at Cornell.[12] Over time, no doubt, the structure and practices of the global digital library will emerge from these and many, many other initiatives addressing various aspects of the hybrid library challenge.[13]

Modes of publication

The big challenges in publication are - rightly - well publicised, and we will return to them, but I'd like to begin by considering the question from the point of view of an individual project, where one of the primary objectives is likely to be to make its information available to as wide a scholarly audience as possible. This is a question that raises technical design issues at a very practical level, and in all the examples we have considered thus far there has been a conscious decision to 'protect' the user from the technology as far as possible.

This approach characterises many projects, and in the projects in which CCH is involved we have articulated the principles as a 'general model' comprising three levels of access, which for convenience we have labelled 'browse', 'query' and 'specialised' access. With 'browse access' materials are presented in standard world wide web format as a set of indices, from which 'point and click' is all that is needed to lead the user to the underlying data displays. The key characteristics of this level are no or minimal technical barriers to use, and minimal prior understanding of the data. With 'query access', one or more search forms are provided, also web-based, allowing users to construct enquiries based on simple or complex criteria. This requires a greater degree of understanding of the data, but still raises no technical barrier. The third level, 'specialised access', is for users who are willing to learn - or already know - the computer language which will enable every ounce of the complexity in the

[11] The Arts and Humanities Data Service (AHDS) is funded by the UK's Joint Information Systems Committee (JISC) and Arts and Humanities Research Board (AHRB). It has a co-ordinating Executive, and five 'Service Providers', covering Texts, History, Architecture, Visual Arts, and Performing Arts (http://www.ahds.ac.uk).

[12] FEDORA is an acronym for 'Flexible Extensible Digital Object Repository Architecture', developed by Carl Lagoze, Sany Payette and colleagues at Cornell. Since the time of the talk, the Andrew W. Mellon Foundation has funded a collaborative project based at Cornell and Virginia, involving Lagoze and Payette at Cornell and Thornton Staples and colleagues in the Digital Library Research and Development Group at Virginia, to develop a practical implementation of the FEDORA model (http://fedora.comm.nsdlib.org/).

[13] See also: Deegan, Marilyn and Tanner, Simon: *Digital Futures: strategies for the information age*, London, 2002.

underlying system to be exploited. In the case of relational database projects, for example, this is likely to be Structured Query Language (SQL).

Part of the reason for this three-tiered approach is to maximise the use and the usefulness of these expensively created resources. It is anticipated that the browse level mechanism will suffice for a substantial number even of scholarly users, as well as opening the resources up to a wide range of non-specialist use, e.g in schools and public libraries. With the addition of the form-based tier, it is expected that almost all scholarly needs will be met. And for a few hardy souls, there will still be the final level!

Returning to the more major issues, I don't propose to say a great deal since they are well rehearsed. The roles and relationships of scholars, libraries and publishers are having to change, and there are no clear models for how to proceed. The economics of publishing are changing, and this has given significant impetus to thoughts of institution- or consortium-based approaches, and the Open Archives Initiative (OAI: http://www.openarchives.org/) and the SPARC initiative, with its slogan "Returning Science to Scientists" (http://www.arl.org/sparc/), may be important in this regard. The new initiative at the University of Virginia to create a department in the University Press with specific responsibility for 'born-digital' resources is also very interesting (http://www.iath.virginia.edu/imprint/).

New methods and new tools

The application of computing is forcing a re-evaluation of research methodology in the humanities, partly because of the scale of evidence these methods make available for systematic analysis, as previously alluded to, and partly because of the rigour demanded by computational processes.

Earlier this year we organised a colloquium at King's entitled 'Humanities computing: formal methods, experimental practice', held on 13 May 2000 (http://www.kcl.ac.uk/humanities/cch/seminar/99-00/seminar_hc.html), which drew on such fields as computer science, sociology, and philosophy and history of science, as well as the humanities and humanities computing. It was an attempt to address issues of inter-disciplinarity, of which more later, and to explore the extent to which it may be useful to think of the application of computing in the humanities as an experimental science.

Scholarly primitives

One paper in the colloquium tackled the question of whether humanities computing should identify, and if necessary create, a set or sets of scholarly primitives, whose combination and re-combination would enable researchers to shape and re-shape their data in the ceaseless quest to find patterns and discontinuities. TUSTEP is a fine model here, with its modularity linked to the what are described as 'the small steps' of progressing through the work. More think-

ing of this kind is needed. The ideal is a set of protocols enabling interchange of data and interoperability between systems.

Modelling and structured design

Inherent in any mark-up of text, the creation of any database, the creation of any digital resource, is an iterative process of analysis and modelling, with the finished work being an instantiation of the final model in that process. At its best the process forces researchers to view their data in a new way, to confront new aspects of inconsistency - and at times to consider whether the data is really susceptible to the rigours of a formal analysis and modelling process. Productive - and necessary - though these methods are at their best, they are not widely known and understood, and their wider dissemination is one of the important challenges we face.

Statistical methods

There use of statistical methods, e.g. in stylistic analysis and authorship attribution studies, is well established. Although not seen as 'glamorous' in comparison with the latest multi-media fashions, the work has been continuing steadily, due to the efforts of Burrows and others, as previously mentioned, and has been making gains as new statistical methods are proposed and tried. In some areas of historical research their use has been much more widespread. It seems clear that for many humanities researchers, the methods continue to seem forbidding, partly because of a reasonable suspicion of attempts to 'measure' the creative imagination, and partly through lack of awareness about the possibilities of statistical methods and lack of training in the associated techniques and procedures. Professor Anthony Kenny, past President of the British Academy, said in the 1991 British Library Research Lecture:

> "The third lesson is that it will not be possible for humanists to take full stock of what the computer has to offer their disciplines until the study of statistics becomes a normal and inescapable part of the training of those who plan an academic career in the humanities. This needs to be recognised not only at university level (as in France, where now a course in statistics is an essential part of the training of an academic historian) but also in any high school which has an interest in sending on students to do university work in the humanities." Kenny, A.: *Computers and the Humanities*. Ninth British Library Research Lecture. London: British Library, 1992, p. 10.

This may well be another of our challenges.

Institutional models for humanities computing

One of the things our two institutions - Tuebingen and King's - share is a particular conception of how best to provide the framework for scholarly work in the humanities using computers. This conception is based firmly on a notion of creative interaction between scholarship and technology.

Let us begin with the LDDV department here in Tuebingen, whose anniversary we are celebrating. It consists of academic staff and programmers, and it is worth quoting from (the English version of) Professor Ott's description of the LDDV:

> "... This implies that service and research are closely related: part of our service ... consists in the research in data processing methods and in the development of tools for computer-aided philological research. E.g. we do not prepare critical editions or carry out historical research projects, but we develop software ... and collaborate with the scholar responsible for a project in designing methods for new applications."

On the web site of my own department, the Centre for Computing in the Humanities (CCH) at King's College London, you will find similar language:

> "The primary objective of the CCH is to foster awareness, understanding and skill in the scholarly applications of computing."

All of us in CCH have both humanities and computing backgrounds, and it would not be possible to do what we do without this.

A third institutional example to which I would direct your attention is John Unsworth's Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia. Its work also is based on the principle of scholarly and technical collaboration. One particularly interesting feature is their programme of research fellowships, which involve support from IATH, a semester of teaching relief and support for a graduate research assistant. You have only to look through the projects undertaken at Virginia under this programme, including the Blake Archive, to be impressed both at the range of scholarly interests and the variety of technical methods they encompass.

My starting point was infrastructure, and I selected three models that share certain features I believe are important. It would be wrong, however, to leave the impression that these are the only insitutions that provide infrastructures along these lines, or that these are the only good models. My colleague Willard McCarty has begun to develop a typology of institutional organisational models

for humanities computing,[14] and with the support of the ALLC I hope this work can be taken forward. The new opportunities for humanities computing make it more important than ever that institutions develop appropriate frameworks to support the scholarly work, and also mechanisms to recogise and reward innovative work in this field. These are among our most important challenges.

National and international frameworks

I have concentrated on institutional matters, but it would be wrong to ignore the significance of national and international activities. The scholarly and professional assocations have an important role to play, with the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities being the major ones in our field. There is also an increasing number of national and international agencies and projects that are relevant, some mentioned earlier in this paper, e.g. the TEI and the UK's Arts and Humanities Data Service.

New partnerships

Working with scientists and engineers

There are many examples of scientists and engineers becoming engaged with and helping to address the problems of research in the humanities, and I'd like to mention just three by way of illustration.

- The first is Kevin Kiernan's work with scientists at his institution, the University of Kentucky - as well as with the manuscript curators at the British Library! - on the Electronic Beowulf project: http://www.uky.edu/%7Ekiernan/eBeowulf/main.htm. The key scientific work was image processing that allowed much finer analysis of fire- and preservation-damanged manuscripts of the medieval English poem *Beowulf*.
- The second is the work with engineers that is being undertaken by Oxford's Centre for the Study of Ancient Documents, in which a number of engineering techniques are brought to bear on a wide variety of texts inscribed on a number of different types of material: http://www.csad.ox.ac.uk/CSAD/Images.html.
- My third reference is to the *OMRAS* project, which involves musicologists and electronic engineers at King's College and the University of Indiana. It is carrying out research on the automated recognition of 'musical objects', research with far-reaching potential within the academy and in the wider commercial and entertainment world: http://www.kcl.ac.uk/kis/schools/hums/music/ttc/IR_projects/OMRAS/.

[14] McCarty, W. and Kirschenbaum, M.: *Humanities computing: institutional models and resources*: http://www.allc.org/hcim

The point about these examples, and many others, is that partnerships of this kind are based on the use of computers, but very specialised use, developed originally for purposes far removed from the humanities. The challenge is to seek out and exploit these new opportunities, and to create environments in which such partnerships arise 'naturally' and can flourish.

The cultural heritage sector

Partnerships with museums, galleries and national libraries are much more familiar to humanities scholars. Yet the rapid advance of multi-media and virtual reality techniques offer many new opportunities, and require new ways of thinking. The creation of 'virtual collections' of physically separated objects, or of objects of different kinds - material artefacts and manuscripts, for example - is just one example.[15]

New modes of collaboration

Earlier I talked about the structures and the approach of your department, my centre, and IATH at Virginia. One of the key features in common between these three institutions is a founding principle of collaboration, of a joint activity to which each participant brings specialised skills and experience, and differing considerations of theory and of practice. It is this concept of collaboration, and the sometimes unexpected consequences as the different sets of theory and practice meet, that characterises the work we do, and makes it a constant source of challenge - and, we hope, reward.

I raise the matter again because apart from the research under way at Virginia mentioned earlier, there has been insufficient effort to understand and document systematically the many facets of these new modes of collaboration. One of the key challenges we face is to understand them better and to train scholars in the issues of principle and practice which characterise them.[16]

Continuity and a culture of change

Continuity of citation is fundamental to our traditions of scholarship, yet there are many uncertainties in the digital age. Some of the most complex issues are those being addressed by the digital library research I have referred to - the development of metadata structures to ensure adequate citation information, of 'persistent' addressing mechanisms, of sophisticated means to manage digital

[15] I note, in passing, that the Blake Archive numbered among its collaborators 8 galleries and a private collector: http://www.blakearchive.org/public/institutions.html.

[16] For an earlier discussion of new modes of collaboration, see Unsworth, J.: *Networked Scholarship: The Effects of Advanced Technology on Research in the Humanities*, in: *Gateways to Knowledge*, ed. Larry Dowler. MIT Press, 1997.

objects, of long-term access and preservation. This remains one of our major challenges.[17]

Digital culture, digital scholarship

It is not only in dealing with the digital representation of 'traditional' source materials that the humanities scholar is having to become 'digitally literate'. We live in a 'wired world' - a strange term, perhaps, to describe what is in reality an increasingly wire*less* world - in which creative activity of many kinds is carried out by digital means, from literary or musical composition to performance art. Our contemporary culture is increasingly 'digital', we find both the old and new subjects of our study appearing in digital form, and the boundaries between the digital and the non-digital become increasingly blurred, never more so than in the world of 'virtual reality'. Perhaps the overarching challenge we face is 'digital scholarship'.

Marilyn Deegan has for many years been working at the forefront of humanities computing and digital library developments. In her keynote address at the ALLC-ACH 2000 conference in Glasgow, entitled 'Digital Scholarship in a wired world', she said:

> "My contention is that digital scholarship in a wired world *is* different in some profound way from the scholarship that has gone before it."
> Keynote address at the joint international conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, University of Glasgow, July 2000. Unpublished; quoted with permission. The manuscript of her paper was invaluable in the preparation of mine, and I acknowledge this with thanks. –
> Deegan is Digital Resources Director of the Refugee Studies Centre, University of Oxford, Editor of *Literary and Linguistic Computing* and Co-Director, with the author, of the Office for Humanities Communication.

This is a challenging proposition, and Deegan, McCarty and I are working to arrange a colloquium - or perhaps a series of colloquia - to pursue it in greater depth.

---

[17] See, for example: Law, Derek G.: *The Mickey Mouse world of humanities scholarship*, in: *DRH99: A Selection of Papers from "Digital Resources in the Humanities 1999"*, Office for Humanities Communication, 2000.

The role of humanities computing

Bridge, glue and intermediary

'Bridge' encompasses the occasions when humanities computing is the means by which scholars and projects know about and use the techniques and technologies most appropriate to their purpose. And a good bridge must have solid foundations on both banks of the stream (or chasm) it crosses!

Humanities computing is an inter-disciplinary endeavour above all. *'Glue'* describes in particular those projects or other scholarly ventures in which it contributes the technical methods that bind together the multiple disciplines and media that make up the whole.

'Intermediary' describes the pro-active role humanities computing has to play in seeking out potential partners and in identifying scholarly activities that can take advantage of the technologies. It should also be monitoring and assessing changes in technology for their potential benefits to humanities scholarship. In the best cases computing humanists will be in a position to influence the technological developments, as with the TEI and XML.

The role of humanities computing must reflect its soul, and at its core it is interdisciplinary and methodological.

Pushing the boundaries

Recent experience has shown how significant is the effect of technological development, especially in computing and communications technologies, in removing or reshaping boundaries. If we are alert and careful, humanities computing is in a unique position to push boundaries in directions that will benefit rather than distort or trivialise humanities scholarship, and to promote the innovative thinking on which its future in a highly competitive and fragmented cultural environment will ultimately depend.

New ways of thinking

Now I want to return to the question of analysis and design, which I mentioned, for example, in relation to the prosopography projects at King's. Observing the historians in the design phase of these projects, it is clear that having to think in such a highly structured mode forced and enabled them to confront their scholarly materials in a new way and this exercise in new thinking gave them new insights.

Transforming the disciplines

Staying with prosopography to introduce my next point, it has also been fascinating to observe at close quarters a discipline in the process of transformation.

Before computers, prosopography involved the reading of sources by 'the prosopographer' and the preparation of scholarly synthesis and summary based on the sources. Computational methods make it possible to record what all the sources say, so that the synthesis and interpretation can be done by any user of the resource. It would be possible, of course, to imagine that this resource would contain not only the structured database but also the full texts of all the sources. One could further imagine these resources being interlinked. In such a circumstance, any scholar would be able to compare the sources and carry out their own synthesis. What kind of new definition do we then need of 'prosopographer' and 'prosopography'?

Similar isses are being raised in relation to textual editions - a subject with a growing literature.[18] It is perhaps inevitable that the new methods are calling into question traditional roles and long-established ways of doing things. This points to one of the key roles of humanities computing, as a mediator in the changing academy.

A new research agenda

At various times and places in the past year or two, Willard McCarty has sought to provoke discussion and thinking on whether 'humanities computing' should be conceived as a discipline. As part of his research and writing on the subject, he has raised the question of whether humanities computing has a research agenda that it could claim as its own. In at least two key areas it seems clear that there *is* readily identifiable and significant research to be done: on the methodologies that are common across a number of humanities disciplines; and on what occurs at the interface between scholarship and technology, and the effects of the interaction on both. There are likely to be other areas: perhaps in new definitions of scholarly and technical primitives and the development of new tools that might grow from this; perhaps in relation to cognitive science and the question of how we know what we know.

It is in developing this distinctive research agenda that humanities computing will perhaps best prepare itself for the significant roles it will continue to have within the broader context of humanities scholarship.

[18] Inter alia, see: Sutherland, Kathryn: *Electronic Text: Investigations in Method and Theory*, Oxford, 1997