

# Computing or Humanities?

## *The Growth and Development of Humanities Computing*

By Martyn Jessop

The application of computing to research problems in the humanities is not new. One of the acknowledged pioneers in this area, Father Robert Busa, began his work on the Index Thomisticus (an index to the works of the medieval theologian Thomas Aquinas) in the late 1040s, very soon after the first stored-program computer was developed. Many followed his lead, including Antonio Zampolli who pioneered the application of computational techniques in literary and linguistic research from the 1960s. Despite this long association of applied computing with humanities research it is only in recent years that the application of computing techniques has become widespread among humanities scholars. The reason for this is almost certainly that it has taken a long time for computing techniques and technology to advance to a state where they can process the myriad and complex data sources of the humanities and begin to answer the manifold questions asked by researchers.

What do we mean by the humanities? Well, a brief list of humanities disciplines might include (among other things), American studies, Anthropology, Archaeology, Art, Art History, Byzantine and modern Greek Studies, Classics, Comparative Literature, Cultural and Creative Industries, English Language and Literature, Film studies, French, German, History, Medieval studies, Music, Palaeography, Philosophy, Theology, Visual and Performing Arts. Scholars in these disciplines are working with source materials that have been produced by human hand and mind in a huge variety of ways. These materials range from ancient inscriptions on stone, through a wealth of written and printed materials, works of art in many different media, cultural artefacts, sound recordings, live performance and digital moving images.

The scholars in these disciplines need sophisticated digitisation techniques and equipment, high capacity storage devices, powerful processors, image processing techniques, high quality display devices and many other products of technological advances that have only recently become available at a reasonable cost. In order to make effective use of computing in their research, scholars also need to apply

knowledge from other fields, especially computer science and information science. The traditional ideas of single discipline knowledge have given way to a more multidisciplinary approach; this in turn has developed further as the level of integration between knowledge from a range of humanities subjects and from many facets of computing and information science has been drawn together into the new discipline of Humanities Computing.

This paper discusses the nature of humanities computing in a pragmatic way by looking at the three components that form any branch of applied computing, namely: the data used by researchers, the computing tools they apply to it, and the knowledge that underpins their research.

### **Humanities data**

Humanities scholars work with the products of the human hand and mind from civilisations that span our entire history. The humanities disciplines are often associated with the study of textual sources, from ancient stone tablets and papyrus that are thousands of years old to the content of emails from contemporary politicians. But textual sources form only a small part of the raw data of humanities scholarship. Much of what is studied is in the form of objects such as sculpture, paintings, archaeological finds, designed objects produced by modern industry, architecture as diverse as Babylonian arches to the Trump tower and so on. Scholars may also wish to study performances as different as Ancient Roman theatre or contemporary dance. These sources need to be represented in digital form before computing techniques can be applied to them. They can be represented digitally in the following forms

Text -- increasingly textual sources are "born digital" but the majority of the sources studied by humanities scholars are in analogue form. Text may be in any ancient or modern language and from any of the variety of media used through the ages. The term "text" may also be extended to other forms of symbolic notation such as music.

Numerical data -- common to many disciplines, humanities scholars rely on numerical and statistical analysis of sources such as historic census data, as well as that derived from textual analysis techniques.

Images -- many of the humanities disciplines, such as art history, are highly visual but even the less visual often rely heavily on objects or materials that are represented in digital images, early manuscripts for instance.

Moving Images -- Film studies make extensive use of digital moving images, but film and video materials are also important to the visual and performing arts. They are also used as recording media and for teaching purposes in disciplines throughout the humanities.

Spatial data -- many of the humanities disciplines deal with maps or data that have a spatial component. The use of the term "spatial" here need not be used in the conventional geographical sense but can refer to the spatial distribution of compositional elements in a painting, the positions of dancers or actors on a stage and many other sources involving location data.

Sound -- sound recordings are of immediate interest to music and performing arts scholars, and to the disciplines where the study of language is a component.

In addition to these types of data it is also necessary to have metadata, quite literally data about data. Metadata performs two key roles in humanities computing. Firstly, as in other fields, it is used to identify and classify digital objects, thus allowing them to be discovered by searches. The second major role of metadata is to render the content of digital objects visible to computing techniques and thus allow scholars to apply both traditional and new research methods to the data. Mark-up languages such SGML and XML have been developed and standards defined by initiatives such as the Text Encoding Initiative (TEI). The developing technologies of Content-Based Image Retrieval (CBIR) will find far ranging applications in the humanities by providing scholars with the ability to discover, organise and research the visual material that underpins a substantial part of humanities scholarship.

The acts of both representing and analysing these research sources in digital form draws upon a large body of knowledge from many facets of computing and information science. This knowledge has to be applied in a way that satisfies the scholarly criteria of each of the humanities disciplines, which have both

commonalities and differences. Humanities computing seeks to integrate knowledge from these huge and diverse fields into a single field of study. The applications we use and develop are applied across the full breadth of humanities subjects but we seek to identify a core that is common to them all.

### **Computing tools for Humanities Research**

Having digitised the source data and organised or annotated it in ways that make the scholarly content visible to computing techniques, we next need to look at the computing tools that will allow scholars to perform research upon them. These tools have been lacking in the past and their development is one of the threads of humanities computing. But what exactly is it that scholars do when they perform research, and why do they need tools to help them do it?

John Unsworth (2000) has suggested that although the sources, subject knowledge and outcomes of research in the humanities are extremely diverse it is possible to identify certain methods they all have in common. These research methods can be expressed as a set of seven scholarly primitives. These low-level methods combine and feed into each other to form the basis for higher-level scholarly activity throughout the humanities. Unsworth defines them as:

- Discovering
- Annotation
- Comparing
- Referring
- Sampling
- Illustrating
- Representing

A primary task of humanities computing is to provide the technological tools to allow academics to apply these primitives to the range of digital data and resources available across computer networks and to ensure the viability of these resources into the future.

*Discovering* is what academics have previously done in archives and libraries etc. They now need to apply this primitive to digital resources on the web and elsewhere. The most obvious example of this are Internet search engines. Another "traditional" method of discovery is via conversations with others and increasingly the medium for this form of communication is the Internet. Online discussion groups have been heavily used for many years by humanists, and communities of scholars who correspond and work together using a mixture of electronic and conventional means have emerged. This has meant that there have been some changes in the kinds of projects that scholars engage in, with collaborative work being more common than hitherto. The growing use of collaborative Websites, wikies, and blogging mark the next phase of development of discovering through communication with others.

*Annotation* has always been an important technique to academics in the humanities, and even in that most traditional of analogue scholarly media, the book, we frequently find personally added marginalia. But how can this ability to add the personal thoughts of a scholar be applied to electronic texts or images, or any of the other digital sources such as sound recordings or moving images? Tools to facilitate annotation for each of the forms of electronic information are under development, and a key issue here is often not how to facilitate annotation but rather how to share these annotations between scholars in a way that is open but also secure from abuse or accidental damage. One of the case study projects described later allows users to annotate bars of music manuscripts online, and share these annotations with others.

*Comparing* is the ability to compare two or more objects of analysis. This is traditionally done visually, for example comparing the same passage of the bible from different versions side by side, but now sophisticated tools for automated comparison of textual data are available that can compare up to 100 versions of a text simultaneously and report the points of difference. Comparison can also involve numerical or statistical analysis, a common example being the comparison of census data from one period with that from another. Techniques for comparing images, sound and moving image data are under development.

A common example of *referring* occurs when operative associations are created between, among and within digital objects. These are often in the form of links between one fragment of information and another, the very essence of the Internet.

An important aspect of reference is the need for stability, and a manifestation of the lack of stability is the ever-present problem of "link rot" on the Web, and the rapidity with which information on the Web is changed, often with no means of tracing the changes. This is a serious problem in the humanities.

*Sampling* is the result of selection according to a specific criterion, which could be a search term or rate of sampling frequency, for example one frame out of every five from a celluloid movie.

*Illustrating* is the process of elucidating or making something clear. This can take many forms from simplifying complex results to diagramming and graphical representation.

*Representing* is most commonly seen in the publishing process whether conventional or electronic. The products of many research projects are now published as digital resources, as well as conventional books and journal articles.

The scholarly primitives described are of course not limited to the humanities, but are research fundamentals in other disciplines. What defines the differences for the humanities is of course the nature of the data, the nature of the questions to be asked, and the tools to be used for the various interactions. The critical research data tends to be primary source materials (books and manuscripts of all periods, historical and archaeological artefacts, art objects etc) as well as the whole range of secondary sources. These are held in cultural institutions of all kinds: libraries, archives, galleries, museums, as well as being part of the landscape (buildings, public monuments, archaeological sites). There have been a large number of digitisation projects over the last 10-15 years that have made a substantial volume of scholarly research materials available. There have also been projects that have developed interfaces, presentational tools, analytical tools, resource discovery tools, as well as projects to develop standards for mark-up, description and long-term preservation of humanities data in digital form.

## **The Methodological Commons**

The scholarly primitives are just the building blocks of humanities computing tools and processes. They do not define what humanities computing actually is. Stepping beyond the data and scholarly primitives, Willard McCarty and Harold Short (2002) have produced a preliminary intellectual map of humanities computing (<http://www.kcl.ac.uk/cch/allc/reports/map/mapframe.html>). The humanities disciplines are at the top of the map organised in groups, such as literary and linguistic studies, historical studies, material culture, musicology and performance studies.

At the bottom of the map are broad areas of learning that interdisciplinary work in humanities computing calls upon: philosophy (areas such as epistemology, ontology and the philosophy of mind), historiography and ethnography, literary criticism, linguistics, science studies, sociology of knowledge, media studies and numerous aspects of computer science, such as mark-up technologies, digital library research and image processing.

At the centre of the map is a large "methodological commons" of computational techniques shared among the disciplines of the humanities and closely related social sciences, e.g., database design, text analysis, numerical analysis, imaging, music information retrieval and communications. Each disciplinary group contributes techniques to the methodological commons. As new applications of these techniques are demonstrated in other disciplines they in turn are exported from the commons into new disciplinary groups. Humanities computing is the agency that oversees this development process, taking methods from one discipline, developing them and then applying them in other disciplines. Part of this process is the identification, or creation, of the tools to fulfil the roles of the scholarly primitives described earlier. The tools do not exist in isolation; they must be developed and used in ways that satisfy the scholarly criteria of all the disciplines involved in their production.

## **Case Study of a Humanities Computing Project at King's College London**

King's College London has shown a high level of commitment to developing the effective use of applied computing in research, teaching and learning in the

humanities disciplines. Their support has allowed the Centre for Computing in the Humanities (CCH) to play a central role in developing humanities computing at King's and in the wider academic and cultural heritage communities. At King's College, CCH performs a dual role as collegial service to the disciplines and as research enterprise directed to investigate evolving methodologies, devise new computational approaches, study the effects, and tease out the implications. There are two aspects to our work in this role; research and teaching. We collaborate with a range of research partners from departments within the college, other institutions and external cultural heritage organisations. The experience gained from research projects in turn enhances other projects and feeds directly into the education of the next generation of scholars.

The following case study is offered as an illustration of the complex nature of humanities computing. The two linked projects described below are close collaborations between the Music Department at Royal Holloway College, University of London and the Centre for Computing in the Humanities (CCH), King's College London. The projects offer enhanced digital access to primary sources of the music of Fryderyk Chopin.

The first project, Chopin's First Editions Online (CFEO), has as its main aims:

- To create an online resource uniting the original impressions of Chopin's first editions in an unprecedented virtual collection (an archive comprising 4,345 digital images of Chopin's first editions will be available online)
- To develop complex textual interlinking of this virtual collection and relevant excerpts of the Annotated Catalogue of Chopin's First Editions
- To provide comparative text-analytical commentary on the multiple first editions in this archive
- To devise innovative technical methodologies for complex Web delivery of this material, using advanced imaging techniques allied with relevant open standards for metadata and interface design.

This project is funded by the UK's Arts and Humanities Research Board

The second and closely associated project (funded by the Andrew W Mellon Foundation) is investigating the use of emerging technical capacities for text/image comparison and new music-recognition technologies for the creation of an Online Chopin Variorum Edition (OCVE). The project deals with a number of data types text, images, sound and metadata, and the content of the images as a form of spatial data, with the location of content in the image being of special importance. The project's primary scholarly goal is to facilitate and enhance comparative analysis of three categories of source material: manuscripts (sketches, autographs, scribal copies, glosses in student copies, etc.); first impressions of the first editions (being digitised and made available through CFEO); and later impressions of the first editions (i.e., those pages of the first editions containing variants, whether attributable to the composer or to others involved in the editorial process).

Referring back to John Unsworth's list of basic tools we can look at the scholarly primitives that can be applied by researchers to the resources made available by these projects. A variety of resources can be "discovered" by different searching mechanisms. The manuscripts can be annotated by the user and different versions compared by juxtaposition or superimposition. Links can be established between different manuscripts and other sources. Sampling will be performed within the searching mechanisms. Images and notes will be generated which will illustrate, and elucidate, the issues that music scholars are researching. Finally the whole project is published on the Web, thereby fulfilling the goal of representation.

The tailor-made use of new technology in this way will enable comparative analyses of disparate types of source material, attaining a level of manipulability far outstripping that of existing printed variorum editions of Chopin's music and indeed of any composer to date. Although the OCVE pilot study will make a specific contribution to Chopin scholarship, it is therefore also intended to pave the way for detailed investigation and analysis of the music of other composers.

The scope of humanities computing is far greater than a single project can show. However the discussion of these projects does give an indication of the style of the end products that the application of humanities computing can produce and shows some of the techniques and technology used. It also demonstrates the range of knowledge that must be drawn together from humanities disciplines, computer

science and information science. This style of project invariably involves specialists from many disciplines working together in a highly collaborative team.

## **Conclusion**

The range of possible uses of ICT in humanities data is now considerable, and includes text analysis; stylistics; advanced mark-up and metadata techniques; electronic editing and stemmatics; multimedia and hypertext theory; electronic intertextuality; formal methods in the modelling of textual data; manuscript studies; paleographic techniques; codicology; multilingual textual work; corpus linguistics; linguistic tagging and analysis; image analysis; modelling of image databases; visualization; advanced image search techniques such as QBIC (Query by Image Content); image enhancement and virtual reconstruction (manuscript images etc); multimedia image databases; 3-D modelling for archaeology, theatre studies, other artefact studies (sculpture, architecture, fine art); layering of archaeological site information; mapping and GIS; structured data techniques including advanced data modelling, data and project design, prosopographical data design; CAD/CAM techniques in art and design; digital art; live art and performance; digital animation; applied art; new modes of publication involving advanced methods, including e-book and e-journal publication, multimedia publication, and advanced techniques in web publication.

The methodological commons described by McCarty and Short defines broad areas of learning that the interdisciplinary work inherent in humanities computing calls upon. Their map also shows the range of disciplines to which the methods within the methodological commons are applied. The commons is continually evolving as new techniques are absorbed from one discipline and ways of applying them in innovative ways to new fields are developed. Humanities computing is the agency that oversees the evolution and development of the content of the methodological commons. It strives to integrate knowledge from huge and diverse fields into a single field of study. In doing so it has to satisfy its own scholarly requirements and those of the disciplines to which it is applied.

As a discipline humanities computing shares the characteristics of many branches of applied computing. It has a constantly evolving core of knowledge and

methodologies. This process of evolution is reflected in our teaching of the next generation of humanities scholars. We develop the students' aptitude to readily and inventively adapt to rapidly changing systems and circumstances. This style of humanities computing teaching and learning aims to instill knowledge from the methodological commons described earlier. The majority of humanities disciplines change relatively slowly but the humanities computing knowledge we teach must be able to survive rapid change and have the ability to respond to the requirements of the moment.

## References

The Centre for Computing in the Humanities, King's College London  
<http://www.kcl.ac.uk/cch> accessed 12/10/2004

Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?. Unsworth. J.  
<http://jefferson.village.virginia.edu/~jmu2m/Kings.5-00/primitives.html> accessed 12/10/2004

The Text Encoding Initiative (TEI) website <http://www.tei-c.org/> accessed 12/10/2004

Mapping the Field, McCarty, W. Short. H  
<http://www.kcl.ac.uk/cch/allc/reports/map/mapping.html> accessed 12/10/2004

Martyn Jessop, Centre for Computing in the Humanities, King's College, London;  
email: [martyn.jessop@kcl.ac.uk](mailto:martyn.jessop@kcl.ac.uk)

*Source: Ubiquity, Volume 5, Issue 41, Dec. 23 - 31, 2004,*  
<http://www.acm.org/ubiquity/>