Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 50 (2004) 261–276
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

CITRIS and data and knowledge engineering: What is old and what is new?

Ruzena Bajcsy [a], Rick McGeer [b,*]

[a] *CITRIS (Center for Information Technology Research in the Interest of Society), University of California, Berkeley, CA 94707-1764, USA*
[b] *Hewlett-Packard Labs, Mail Stop 1167, 1501 Page Mill Road, Palo Alto, CA 94304, USA*

Available online 4 May 2004

Abstract

We survey the issues involved in knowledge capture and representation for the humanities and outline the design of a modern information system, which will permit humanities researchers to easily publish, share, view and process data. This system takes advantage of recent advances in the capture, representation, indexing and search of text and media data, including representations of three-dimensional images and three-dimensional enhancements of two-dimensional images. We suggest the exploitation of modern networking technologies to make electronic publication accessible to non-technologists. Finally, we discuss the problem of privacy in data mining and outline possible solutions from the recent literature.
© 2004 Elsevier B.V. All rights reserved.

Keywords: CITRIS; Database Management; Humanities; Solving societal scale problems; Data mining; Privacy

1. Introduction and motivation

The Center for Information Technology Research in the Interest of Society (CITRIS) is one of four institutes established by the state of California to promote innovation, which will provide/enhance the quality of life and the prosperity of Californians. The conditions under which these institutes were established were that they be multi-disciplinary and multi-campus while advancing the study of life sciences, information technology, nanoscience and nanotechnology.

* Corresponding author. Tel.: +1-650-857-1501; fax: +1-650-857-4012.
E-mail addresses: bajcsy@eecs.berkeley.edu (R. Bajcsy), rick.mc-geer@hp.com (R. McGeer).

0169-023X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2004.04.002

The CITRIS mission is to investigate the uses of information technology in solving societal-scale problems. While new technology development and understanding of systems is anticipated as a result of these investigations, the primary product of the institute is expected to be applications which have an immediate or medium-range impact on many of society's most pressing problems.

Furthermore we are concerned with problems of applications of Information Technology (IT) on a scale (societal scale) that are typically not explored in academic groups or laboratories. While this applications agenda is broad (transportation, health care, environmental monitoring, energy conservation, disaster mitigation and education), in this paper we will concentrate on applications that serve our cultural and educational institutions (schools, museums, etc.), as well as our policy makers and those who study cultural evolution and social interactions between people and societies. We will argue that the new digital media, data and knowledge engineering, will have a profound impact on understanding, researching and teaching history, archeology, anthropology, political science, sociology, urban development and economics.

2. Knowledge management for humanities as a tool for discovery

Social science and humanities research is increasingly advanced by inter-disciplinary and inter-institutional collaborations. This is largely because human interactions and artifacts, which are the focus of humanities and social science research, are governed by the broad tapestry of human affairs. This tapestry is woven by the threads of institutional disciplines. The triumphs of the Roman legions under Aurelius, for example, were at least in part due to the rise of Stoicism as an influential philosophy and system of beliefs among the Roman warrior class. It has long been known that artifacts produced by societies are strongly influenced by the dominant religion and governing class or empire. Similarly, examination of artifacts can be used to trace the spread of ideas, trade and religions throughout the world. In each of these examples, history, archaeology, anthropology and ancient sociologies meet and no single discipline can unlock these mysteries. It is, thus, imperative that disciplines share data and discoveries in a format common to all, while preserving the critical information of each discipline.

The Internet forms a unique forum for the enablement of such collaboration. But while early adopters of information technology have embraced the Internet as a powerful medium for collaboration, social scientists and humanists have lagged behind. In part, this is due to the natural division between technophiles (astronomers, particle physicists, engineers and business people) and technophobes (literati, humanists, and social scientists). However, in large part this division is because the information systems we developed have been directed toward the technophilic groups named above.

Relational Data Base Management Systems (DBMSs) are designed to facilitate processing, storage and analysis of financial and business transactions. Computer architectures and programming languages are designed to facilitate rapid, large-scale processing of numeric data, which are useful principally to particle physicists, astronomers, and engineers. Modern graphics systems enable many applications, but were designed first around computer-aided design applications in mechanical and electronic engineering and still retain their roots.

Finally, the killer application for humanities researchers, as noted above, is the publication and sharing of data in a mutually comprehensible fashion. Internet-based publication systems still

retain their server roots: design and maintenance of a simple web page requires some technological sophistication, access to privileged resources (space on a web server) and knowledge of an arcane set of tools from web design programs to the structure of Unix file systems. Further, this does not permit the rich sharing of data, but simply more efficient publication of traditional, paper-based scholarship. The web saves scholars trips to the library: it does not by itself, fundamentally *enable* new research.

Contrast this to the situation in the physical sciences. Project BaBar at SLAC has generated almost a petabyte of data on particle collisions and continues to generate hundreds of terabytes per annum (http://www.slac.stanford.edu/BFROOT/www/Public/Computing/Databases/index.shtm). Every day, hundreds or thousands of physics graduate students the world over access the BaBar data to seek new physics in interactions the BaBar detector has captured. This is a fundamental shift in physics research. Previously, only researchers onsite at the detector facility had access to the raw data of collisions whence new particles are born and physics enriched. Examples abound in the other physical sciences as well. The Virtual Observatory at Johns Hopkins University (http://research.microsoft.com/~gray/JimGrayResearch.htm) has stored thousands of stellar images [1], largely derived from the Sloan Digital Sky Survey. Every day, astronomers around the world access the observatory to test new hypotheses about the evolution and structure of our galaxy and universe.

Recent advances in information technology and the collaborative nature of the Internet make it possible for us to provide this same level of advancement to humanities researchers. These advances come in three parts: new peer-to-peer tools, which lower the barrier to publication for humanities researchers; a new understanding of a common *lingua franca* among humanities disciplines; and new tools for the creation and visualization of images, media, solid artifacts and virtual reality experiences. We examine these advances and their potential impact on humanities research below.

3. Issues in data management for the humanities

Data management for the humanities is a much more complex problem than that confronted by most business or scientific applications, the traditional data management applications for the sciences. Consider two traditional scientific applications: accelerator detector events and the Virtual Observatory. In both these cases, the data to be stored is quite large, in the petabyte range. However, the number of data sources in both cases is small (in the case of accelerator detector events, only the events from a specific detector, such as BaBar or Alice, are stored), in the case of the virtual observatory, only information from a number of sky surveys such as the Sloan Digital Sky Survey, are taken [2]. Further, the data is relatively homogenous and comprises only a few known types. The metadata, or indexing mechanisms, are agreed upon and well known to the practitioners. Finally, there are typically no restrictions on the publication or use of the data, so protection and privacy issues are not confronted in these applications.

Contrast this situation with that of humanities data. The data sets, though smaller, are heterogeneous. Required data types include tabular data, text, images, images with surface enhancements, video, sound, and voxellized three-dimensional renderings of artifacts, monuments, and cities. The data sources are diffuse, varied, and for the most part, without technical support. Similarly, indexing and metadata mechanisms and schemes are highly varied, though there are

basic standards such as the Dublin Core (http://www.dublincore.org). Finally, some of the data—those which reveal personal information about living individuals—have significant privacy restrictions and concerns. The privacy problem here is particularly complex. This will be discussed further in Section 6.3.

Traditionally, the problem of storing, representing and indexing test and image data is the province of digital libraries.

4. Digital libraries, text, images and music

4.1. Digital libraries—text

Digital Libraries have been quite active for more than fifteen years. They also have a great deal of financial support from both federal government agencies (NSF, NIH, DOE and NASA), as well as private foundations, such as the Hewlett Foundation, Packard Foundation, Getty Foundation and many others. As a result, a great many collections have been digitized and archived.

Most universities and cultural organizations have catalogues on line. We list a few examples of successful Digital Libraries:

- California Digital Library (www.cdlib.org).
- The Perseus Digital Library at Tufts University specializing in Classics, Archeology, English Literature, etc. (www.perseus.tufts.edu).
- Alexandria Digital Library (www.alexandria.ucsb.edu).
- The Library of Congress (www.loc.gov).
- The MIT Library (http://libraries.mit.edu/).

Most of these libraries are organized in the Digital Library Federation, which tries to promote the agenda of this area, or other groups such as the DSpace federation (http://www.dspace.org/).

What are the problems? The largest obstacle for making the collections complete is the copyright clause. The result is that most of the digital collections are old manuscripts where the copyright has expired or never existed. The more current journals available on line have negotiated the copyrights and the readers must pay in order to access them. This problem is particularly acute when (as is the common case) an institution's library has paid the access fee on behalf of all the institution's users. The library naturally wishes to make its content available over the network; the problem is then how to prevent access by individuals who are not among those for whom the library has obtained the rights. The general method is to restrict access to requests from IP addresses representing campus users. This fails when an authorized user connects through a third-party connection (e.g., using a cable modem from home, or access through a public hotspot) or when an unauthorized user gains an authorized proxy to obtain the information for him. This was a significant early problem for the CoDeeN proxy network [3]. Another obstacle is the digitization process and more importantly the annotation process, which are still very labor intensive. The standard annotation is by content and author, though recently, more and more collections are also annotated by place and time.

4.2. Digital libraries of images and music

Digitization technology has advanced sufficiently to the point that it is relatively easy to digitize, as well as to store pictures and music. Again many museums, galleries and universities have done so. A few examples being:

- The British Museum (www.thebritishmuseum.ac.uk).
- The Smithsonian (http://www.si.edu).
- The New York Museum of Modern Art (www.moma.org).
- San Francisco Museum of Modern Art (www.sfmoma.org).
- American Museum of Natural History's Digital Library Program (library.amnh.org/diglib).

The open issue is, however, how to access this information. Of course one way is to annotate the pictures by authors, or genre (landscape vs. portraits, etc.). Research is ongoing to extract some pictorial features such as color, shape and texture for indexing purposes [4,5]. Similarly music can be annotated by content, author or by features similar to the pictures. The common concern with this data is authentication and copyrighting. Watermarking algorithms help protect the data, but concern about misuse of this information is still there.

4.3. Three-dimensional capture and representation

Humanities scholars, especially in the areas of cultural preservation, archaeology, and anthropology are concerned with the capture and representation of three-dimensional, real-world objects, of sizes ranging from the small (primitive tools, pottery, statuary, furniture, etc.) to the large (burial mounds, ancient buildings, sites, etc.). New capture technologies, such as laser scanners, multiple small cameras, ultrasound, sensor networks, and GPS-based locators permit the easy capture of once expensive, hard-to-create representations. A comprehensive review range imaging in archeology can be found in Godin and coworker [6]. There are several commercially available digitizing technologies and they work on two different principles:

- One uses projecting a structured light, camera and the triangulation method. This is represented by companies such as Cyberware (www.cvberware.com/products) in larger products and Minolta (www.minolta.com) for smaller objects, like museum size pieces.
- The other technology uses time of flight of the projected light on the surface of the digitized object plus a video camera. This method is employed by companies like CYRA and Delta-Sphere-3000 (info@3rdtech.com). Both of these companies aim at digitizing large objects such as architectural sites. The advantage of the system with time of flight is that is not as sensitive to albedo of the surface as is the triangulation method.

A good review of range sensors development is presented in Blais [7]. Most of the currents systems will provide a complete set of three-dimensional data (x, y, z) plus the color image of the surfaces as well as a mesh.

Imaged artifacts can range from ancient cities and sacred sites, such as Machu Picchu in Peru, large cultural artifacts, such as the temples of Olympia or the bays of Rome's Coliseum [8],

to small artifacts, such as the ancient cuneiform tablets of Iraq. Digitization of this data makes it available to humanities scholars across the world. Coupled with three-dimensional rendering and extrapolation technologies, this digitization promises a world of *Virtual Archaeology*, where discoveries of the ancient world can be made at a computer as easily as in the field. Excellent discussions of these approaches can be found in a paper by Addison and Gaiani [9].

There is clearly an increased interest among both archeologists and computer scientists in the use of IT to enhance computerized methods to rescue Archeology, manage cultural heritage data and provide many other applications. A good example of this effort is illustrated in the program of the Workshop on Archeology and Computer, at the Congress on The E-way into the Four Dimensions of Cultural Heritage given April 8–12, 2003 in Vienna, Austria (http://www.archaeologie-wien.at/).

4.4. Digital libraries of three dimensional visual data

Finally, we would like to elaborate on the opportunity for digitizing, archiving and browsing of three-dimensional spatial data with the possibility of recording the temporal aspects of change for this data. There are several possible applications:

- Architecture and design of both the interior and exterior of buildings and eventually urban sites.
- Objects of cultural heritage stored in museums and/or in private collections (www.hitl.washington.edu/projects/knowledge_base/museum.html).
- Sculptures both old and new.
- Archeological sites.
- Digitization of the human body (for example, the NIH sponsored Visible Human project [http://www.nlm.nih.gov/pubs/factsheets/visible_human.html] or the activities of a person or an entire group of people.

A particularly compelling example of the possibilities of this technology is already available in preliminary form on the web, with the *Rome Reborn* project at UCLA (http://www.aud.ucla.edu/~favro/rome-reborn). Currently available on the web are only the screen shots from the virtual reality tour of ancient Rome that is the focus of the UCLA project; however, nascent web technologies give promise to make this sort of tour available over the web in the near future.

This is a great challenge for Information Technology in several areas: data acquisition, data representation, data interpretation, data visualization and 3D data acquisition.

In the previous section, we covered the available 3D data acquisition hardware. Here we would like to review some algorithmic issues related to data acquisition. The first problem is caused by concave and occluded surfaces, which need to be imaged, independent of whether the acquisition system uses time of flight or structured light—both of which present a problem. The solution is to recognize those locations, which present this problem and position the acquisition system with respect to the concavity, so that the light can reach its surface. This was implemented by Pitto in 1996 [10] for small objects. For larger objects see

Levoy and coworker [11] and Bernardini and Rushmeier [12]. The occluded surfaces can be imaged by rotating either the object or the scanning device. After this step, one has to have an algorithm to patch together the individual views into a coherent data representation. Once we obtain the complete cloud of points, typically a mesh of planar surfaces is generated.

Another technique enhances surface features of objects through polynomial texture mapping [13]. In this technique, a set of images of a particular surface is taken from a variety of known angles. A texture map is then taken which enhances surface textures, allowing for enhancements of small deviations from a smooth surface. The end result is that hidden three-dimensional information, such as eroded cuneiform tablets, fingerprints on surfaces, or even indentations made by a pen or pencil writing on a pad, can be recovered. The total space taken by a PTM is roughly twice that of a conventional image of the same surface.

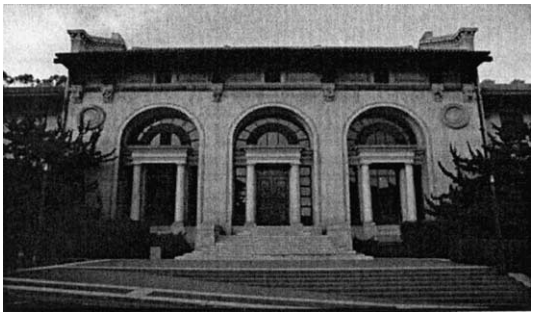
4.4.1. Three-dimensional visualization of data

Coupled with three-dimensional capture is three-dimensional rendering and extrapolation, by which an object can be viewed and (better, in the case of an incomplete artifact) extrapolated through the use of modern graphics technologies. Three-dimensional viewing and rendering is now encapsulated in a widely distributed set of standards, such as OpenGL, (http://www.opengl.org). Further, common tools, such as the embedding of Alice in the Squeak environment (http://www.squeak.org), and visual manipulation of 3D objects [14] bring the manipulation and extrapolation of 3D data to the humanist.

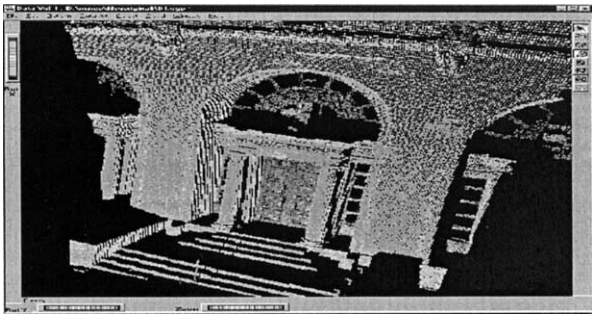
The standard way of visualization is the use of graphics tools for projecting the 3D data from any view the viewer desires. More sophisticated systems provide two stereo pairs of projected data in different colors. The viewer then uses glasses, which fuse the data for the three dimensional appearance.

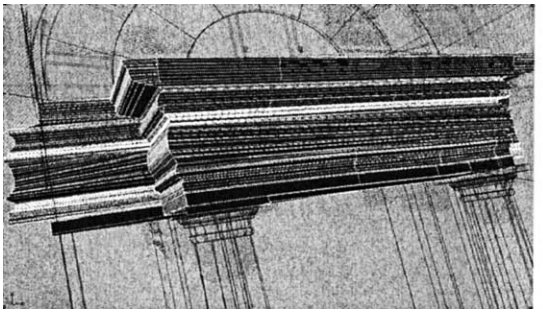
An expensive way of displaying large sets of data is the so called CAVE [15] (http://www.evl.uic.edu/research/res_project.php3?indi=161), which is an enclosed room with walls covered with a continuous display so the user can be immersed in the data with or without stereo capabilities. The intermediate way is to have glasses for the users and pipe the images to them, hence create an immersive feeling for the viewer.

High density scans on UC Berkeley architectural gem, Hearst Memorial Mining Bldg.



Pictures courtesy of Alonzo Addison (UC Berkeley), Marco Gaiani and Federico Uccelli (University of Ferrara) and Daniel Chudak (Cyra Technologies).





Points to CAD to Rendering of UC Berkeley Hearst Memorial Mining Bldg.

4.5. Three-dimensional processing and interpretation of the data, i.e. creating knowledge base

The data processing and interpretation very much depends on the task and application. For example for architects, the data can be processed using computer aided design (CAD) tools, fitting the data to large contiguous surfaces, characterizing the surfaces with some algebraic or geometric functions, dissecting the data according to some functionality, and so on. After such processing, one possibility is to characterize and associate certain styles with some of these mathematical surface/volume descriptions (for example the gothic style, vs. modern or renaissance style.) Another possibility is to compare the evolution of one style over time by measuring the changes in these geometric descriptions.

During the last twenty years, the computer vision community has developed and explored several different representations, both based on surface and volumetric primitives Which of these representations should be deployed will very much depend on the users and queries they will want to pose.

A very different approach will be needed for describing and interpreting anatomical data of humans. As has been shown through the NIH project of the Visible Human [http://www.nlm.nih.gov/pubs/factsheets/visible_human.html], one can virtually dissect different organs, different sections on the body across several organs, or show the overall vascular system, etc.

Longitudinal studies exist showing the anatomy of growth from embryos, through childhood, adulthood and into old age. One can use geometric transformations to measure the changes in growth of the different parts of body over time and quantitatively analyze these changes. Similar methodology can be applied for comparative studies of the physical anthropology of different species, such as Kappelman's e-Skeletons project (http://www.eskeletons.org/).

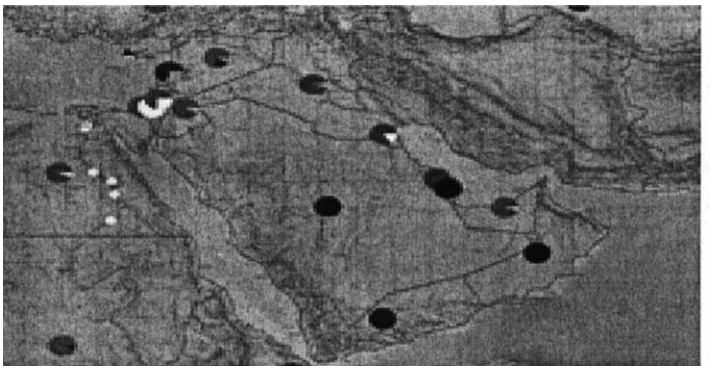
With the current technology of Functional Magnetic Resonance and Positron Emission Tomographic devices, one can superimpose the functional changes over the morphological changes and for the first time, make inferences between the morphological change, functional change and the behavioral change of the subject or subjects. Thanks to the support of the National Institute of Health, there is a wealth of information in the National Library of Medicine on scanning devices and their applications. It also holds much information on quantitative analysis/algorithms for evaluating the images from both the normal population and abnormal populations, such as Alzheimer's, schizophrenia, etc.

We have shown only two different tasks and applications, but the underlying idea is that with the available digitization capabilities and proper geometric analysis, one can shed light on the evolution of natural biological systems over time under varied environmental and social conditions. Similarly, in the case of architecture, with similar analysis one can infer the social, political and environmental influences on society, which enabled the creation of these artifacts.

The problem of indexing metadata is expected to be significant for humanities information systems. Metadata, specifically indexing terms, form the means by which data can be searched and retrieved. In scientific systems the metadata is quite straightforward: the virtual observatory, for example, searches by magnitude and by the coordinates and size of the stellar field. Humanities data is not nearly so easily categorized. Moreover, categorization is not constant from field to field, a severe limitation when data from separate fields are to be compared and used together. Nonetheless, progress has been made. A specific project worthy of note is the Dublin Core (http://dublincore.org/), which forms the basis of all metadata in the humanities.

The above-mentioned analysis will also lead to a possible representation and consequent indexing/annotation scheme for accessing and browsing these complex data sets. Obviously the geometry, parts and their relationships to the whole, will not be the only index for accessing the data. Geo-reference and time-reference will be as important as the content, authorship, political background and other such descriptors. Geotemporal referencing is a very useful representation, especially for humanists and environmentalists, or anyone who deals with objects that are related to location and time. In recognition of this fact, there has been continuous development of Geographic Information Systems and efforts to standardize, coordinate geographic data acquisition and access (http://www.fgdc.gov/metadata/) over the past twenty years. In these systems each piece of data is contextually tagged with a geographic reference, typically expressed as a latitude/longitude coordinate and altitude. For medical applications the Atlas has also a very important meaning because it annotates different structures of the biological system. Examples of Anatomy browsers are plentiful (http://splweb.bwh.harvard.edu/).

Since humanities data is typically highly time-sensitive, these systems are augmented for humanities and social science data with the time coordinate. Such a system is a Geotemporal Information System (GTIS). The most popular common such system is the Electronic Cultural Atlas Initiative (ECAI).



ECAI Dynamic Map
Containing Information on
Iraq and Its Neighbors

GTIS systems provide natural repositories for humanities data since they provide a common lingua franca for all humanities and social science scholars. It is often said that journalism is the first draft of history and many would argue that it is also the rough draft of sociology, political studies and anthropology. That being the case, it is natural that the social sciences concern themselves with journalism's classic five W's: Who, What, When, Where, and Why. Deducing the "why" is scholarly research; "who" and "what" are subject-dependent. But "when" and "where" are truly interdisciplinary, a common tableaux on which humanities data can be displayed and a common basis in the search for objects and events.

The above-mentioned analysis will also lead to a possible indexing/annotation scheme for accessing and browsing these complex data sets. Obviously, the geometry, the parts and their relationships to the whole will not be the only index for accessing the data. Geo reference and time reference will be as important as the content, authorship, political background and other such descriptors.

4.6. Integrated systems

It is obvious that the above technologies exist, or at least are under development, because real applications need an integrated system. What we mean is that there is a need for a unified platform that would enable the user to access information from the digital library of multimedia nature. That is, text, images, sound/music and three-dimensional data of small and/or large objects indexed by content, authors and geo-temporal reference.

The work of Raj Reddy and Michael Shamos from Carnegie Mellon University comes closest to this idea in their project "Universal Library". The Universal Library, as explained in Science, 281, (5384): 1784–1786, 18 September, 1998 is an amalgamation of all recorded human knowledge, searchable from your personal computer. They acknowledge, however, that there is no such integrated system available today. Their estimate is that it will take twenty to fifty years to accomplish this dream. The challenge here is not only the integration of different data formats by making them consistent, but also dealing with the geo-temporal references, especially when one has different time scales, such as seconds, hours, days, years and centuries.

4.7. Open issues

Besides the issues already mentioned in previous sections, such as 3D data acquisition, processing, indexing and display, the open problem is how the users will utilize the system. This is the

standard criticism we often hear. Hence, CITRIS is collaborating with educators in humanities and social sciences on this subject. The Center for Studies in Higher Education at the University of California, Berkeley, under a grant from Hewlett Foundation, is investigating the use of an open knowledge digital collection in the humanities and social sciences in a variety of undergraduate teaching environments, including those in community colleges. The foci of this study headed by Principal Investigator, Diana Harley are enumerated below:

- A preliminary survey of the current use of select open knowledge collections in humanities and social sciences (H/SS).
- Testing the efficacy of a variety of methods to assess actual use of local H/SS collections.
- Understanding the University of California and community college faculty attitudes about their use or non-use of open knowledge collections in H/SS teaching.
- Assembling a cohort of open knowledge owners and digital collection evaluation experts to discuss best practice in assessing open knowledge collection in teaching.

Similar efforts are needed in other areas to test what is desirable and effective in the current technology for the humanist's research and teaching and what advances are needed. The technology is given the necessary resources. The challenge is to collaborate with the humanists who will determine the needed representation, visualization and computation, which in turn can revolutionize their respective fields.

5. Information, communication and search engines

5.1. Peer-to-peer information systems

The rise of peer-to-peer information systems in the late 1990's and early 2000's (Gnutella (http://www.gnutella.com), KaZaa (http://www.kazaa.com), FreeNet [16] offered a fundamentally new mechanism for publishing data to the world. While public attention and debate focused primarily on their intellectual property implications and effect on the business models of the entertainment industry, this obscured a vastly more important reality.

The teenagers who shared their music collections with anonymous strangers had discovered a simple mechanism for publishing data to the entire world. The fact that the data they published was mostly Metallica CDs was incidental. If a 19-year-old can inflict his favorite *Aerosmith* tunes on the world using Gnutella, a Mideast expert could share his translations of Mesopotamian cuneiform tablets using the same mechanism—without knowledge of web publishing or server technologies. Nor would he need funding or permission from a system administrator to do this. He could publish as easily as saving a file. Moreover, the Gnutella model offered a mechanism for distributed, peer-to-peer based search. The protocol did not, of course, specify the search conditions; rather, it permits an implementing application to define a set of search queries and primitives. Similar comments apply to more recent peer-to-peer routing and search projects, such as CAN [17]. Peer-to-peer networking has become a central focus of systems research, which now encompasses large planetary-scale peer-to-peer services networks, such as PlanetLab [18], http://www.planet-lab.org/.

5.2. Search engines

Today's commercial search engines are not designed to find image or sound based information. The problem is that the information on the web is so heterogeneous in structure as to be almost unsearchable using any one approach. There are several ongoing efforts to remedy this, such as Steve Cousins' at Xerox's Palo Alto Research Center and Stanford University [19] which provides a simpler user interface hiding translation systems that can reformulate search requests into many different forms capable of interfacing with different data structures at a contextual level rather than simply matching words. Marti Hearst, of UC Berkeley, has developed Cha-Cha, a program that determines the home page for each item retrieved, records the shortest path to get from that home page to the retrieved page (cha-cha.berkeley.edu).

Searching images, maps, artwork, photographs and video is a more challenging task. Efforts in this area include those of Shih-Fu Chang, an expert in content-based image processing at Columbia University. They developed VisualSEEK, a collection of three search engines that query online art museum collections based on subject matter, physical properties and their like [20], In the UCB Digital Library Project [21], heuristic object recognition was used to index and then search data. It has more recently moved to browsing (see http://elib.cs.berkeley.edu/kobus/famsf/model 2/text and blobs/bbox.html).

Further, search by standard metadata, notably the library standard, has been used and is a feature of the Cheshire search engine [22].

Use of these systems requires programming by subject matter, not technical experts. What is required is an easy, intuitive, natural interface to a GTIS, three-dimensional rendering system with P2P publication capabilities. The *Squeak* environment (http://www.squeak.org), pioneered by Alan Kay (now of HP Labs) is such an environment. In the Squeak environment, end-user programming is principally done by attaching behaviors to physical objects through the use of tiles. This system has been used successfully for end-user programming for seven years and has attracted a wide and loyal user base. What is required is to adapt it to the world of the humanities research by constructing objects for the base-level GTIS data.

5.3. Privacy protection

Humanities data can be broadly split into two parts: data concerning the past, and data concerning the approximate present. The communities that are concerned with each type of data are also largely distinct, though there is some overlap. Archaeologists, historians, and anthropologists are largely concerned with the world of the past. Economists, political scientists, and sociologists are largely concerned with the present world. From the perspective of data collection and presentation, the latter have a wealth of data while the former have a relative paucity. The great problem for those concerned with the present is that the data concerns living humans, and as a result, privacy concerns are immediate and omnipresent. The challenge for those wishing to present data for present-day humanists and social scientists is to preserve the privacy of the subjects while permitting the harvesting of knowledge from the agglomerated collection.

Privacy concerns have achieved statutory status in US census data, and this has hampered scholarly use of this data. The critical problems facing American society can only be tackled if they are adequately analyzed and this analysis can only be done if there is broad access to the

data. However, census and other data are collected under an implicit or explicit guarantee of privacy. Therefore, analysis must respect the privacy of the individuals.

A possible solution comes from the following observation: social scientists are interested only in *aggregate* behavior, not in the behavior of individuals. As a result, data on any individual need not be exposed, so long as data in the aggregate is reliably exposed.

Social science inquiries typically look for correlations between variables in a population: say, between income, education, and voting preference. It is possible, given actual data, to introduce random perturbations to the data to hide the identity of any specific individual, while faithfully preserving all aggregate characteristics, including correlations, in the limit, it is possible to create a population of fictitious individuals with the characteristics of real individuals, respecting the statistical distribution of each characteristic and the correlations between variables. Social science inquiries can be run against this fictitious database, without exposing and sensitive data. Confirmatory tests can then be run against the real database. Unfortunately, it cannot be determined in advance that the database mirrors the behavior of the population database for all *combinations* of variables (in other words, that it respects all n-way correlations). Indeed, statistical data mining of a population is designed to discover hidden correlations between sets of variables that are both previously unknown and surprising. If all n-way correlates had been run, this would amount to running every possible social science experiment on the data. One possible approach to this problem is in the work of Vaidya and Clifton [23], which preserves clustering of data about an arbitrary number of means while preserving the privacy of individuals.

As might be expected, this is an area of active research in the data mining community. Representative approaches may be found in Refs. [23–27]. The goal of these approaches is to determine whether a decision tree procedure, or a classifier, has valid results on perturbed data. Initial statistical results are quite promising and these approaches promise to be the kernel of a privacy protocol for access to sensitive data. Design of the privacy protocol is a large task. First, the programming environment for the perturbed and actual databases must be identical and certified to a level of precision of at least that of the PlanetLab environment. No leaks or unauthorized transmission from the actual database can be permitted. The perturbed data must be certified as sufficiently perturbed so as not to violate the privacy of any specific individual. Finally, error values must be reported to the individual researchers so that the inaccuracies in using perturbed data are faithfully represented.

If a privacy-preserving data-mining proposal can be perfected, the implications for many fields are significant. For example, indications of disease and treatments can be mined from large-scale medical databases. The observation that the coincident rate of Alzheimer's disease and rheumatoid arthritis was extremely low was the first significant clinical evidence that long-term use of non-steroidal anti-inflammatory agents could prevent the onset of Alzheimer's disease. However, data mining of medical records has long been impeded by obvious, valid, and significant privacy concerns. Solution of this in the (less-sensitive) world of social science data therefore has potential to advance medical research.

6. Conclusion

An attractive choice for a knowledge system for the humanities is a peer-to-peer based geo-temporal information system with text, image, 3D and program data all as first class objects. In

this system, each object is linked with common metadata that should include, at a minimum, latitude, longitude, and temporal coordinates. Rather than using a centralized clearinghouse approach, such as that of ECAI, such a system would use a decentralized peer-to-peer approach based on the architecture of Tapestry/OceanStore [28,29] and deploy this as a persistent service on PlanetLab. To the extent possible, this system would use extant open-source systems for its various functions. One example is to use Tapestry/OceanStore as the base layer for object storage and routing through the peer-to-peer network, and DSpace as the persistence/archival layer. Specific subprojects would include:

- Definition of an XML standard for the interchange humanities data, such that this can serve as a substrate for *all* humanities metadata. This standard will include, at a minimum, all fields required by ECAI, including latitude, longitude, time and the Dublin Core. The purpose of this standard is the interchange and indexing of humanities data.
- Creation of a set of content-generation tools to convert existing metadata (web pages, etc.) to the standard.
- Definition of query and indexing protocol, based on the Gnutella 0.4 standard, for the query and exchange of humanities data through the network.
- Creation of a GTIS-based, open-source indexing and query scheme for humanities data.
- Creation of an open-source GTIS-based browser.
- Creation of a set of tools and standards for *gazeteers*, a mechanism by which place names are mapped to geotemporal coordinates.

7. Future work and potential

Just as GTIS P2P coordinate systems propose to revolutionize the study of ancient societies, we expect them to have applications just as important in studies of the current world: from economics to environmental studies to homeland defense to political theory to journalism. When humans interact, where and when are fundamental questions, and their answers provide critical information. Soon, GTIS browsing will be to human- and earth-centered data what Google and other popular search engines are to text.

Acknowledgements

This work is partially supported by funding from CITRIS and National Science Foundation grants CCR-0225610 and CCR 017238-003.

References

[1] A.S. Szalay, P. Kunszt, A. Thakar, J. Gray, Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey.
[2] D.G. York, J. Adelman, J.E. Anderson, et al., AJ 120 (2000) 1579.

[3] V.S. Pai, L. Wang, K.S. Park, R. Pang, L. Peterson, The Dark Side of the Web: An Open Proxy's View, Hot-Nets II or Available from <http://www.cs.princeton.edu/~vivek/hotnets03/>.
[4] D. Forsyth, J. Malik, R. Wilensky, Searching for Digital Pictures, Scientific American, 1997, pp. 88–93.
[5] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, J. Malik, Blobworld: a system for region-based image indexing and retrieval. Third International Conference Visual Information Systems, June 1999.
[6] L. Borgeat, P.-A. Fortin, G. Godin, A fast hybrid geomorphing LOD scheme, SIGGRAPH sketches and applications, 27–31 July 2003, SanDiego, California.
[7] F. Blais, A Review of 20 Years of Ranges Sensor Development, Videometrics VII, in: Proceedings of SPIE-IS and T Electronic Imaging, SPIEVolume 5013, 2003, pp. 62–76, NRC 44965.
[8] K. Moltenbrey, Preserving the past, Computer Graphics, September 2001.
[9] A.C. Addison, M. Gaiani, Virtualized architectural heritage: new tools and techniques, IEEE Multimedia, April–June 2000.
[10] R. Pito, A solution to the next best view problem for automated CAD model acquisition of free form objects using range cameras, in: SPIE Symposium on Intelligent Systems and Advanced Manufacturing, Philadelphia, October 1995.
[11] S. Rusinkiewicz, M. Levoy, Streaming QSplat: a viewer for networked visualization of large, dense models. in: Proceedings of the 2001 Symposium on 3D Graphics, Available from <http://citeseer.nj.nec.com/rusinkiewicz-01streaming.html>.
[12] F. Bernardini, H. Rushmeier, The 3D model acquisition pipeline, Computer Graphics Forum 21 (2) (2002).
[13] T. Malzbender, D. Gelb, H. Wolters, Polynomial texture maps, Proceedings of ACM Siggraph, 2001.
[14] R.W. Bukowski, C.H. Séquin, The FireWalk System: fire modeling in interactive virtual environments, in: Proceedings of the 2nd International Conference on Fire Research and Engineering, Gaithersburg, MD, August 1997, pp. 72–83.
[15] C. Cruz-Neira, D. Sandin, T. DeFanti, Virtual reality: the design and implementation of the CAVE, in: Proceedings of SIGGRAPH 93 Computer Graphics Conference 08/01/1993, ACM SIGGRAPH, pp. 135–142.
[16] I. Clarke, O. Sandberg, B. Wiley, T.W. Hong, Freenet: a distributed anonymous information storage and retrieval system, in: Proceedings of the ICSI Workshop on Design Issues in Anonymity and Unobservability, International Computer Science Institute, Berkeley, CA, 2000.
[17] S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker, A scalable content-addressable network, SIGCOMM2001.
[18] L. Peterson, T. Anderson, D. Culler, T. Roscoe, A blueprint for introducing disruptive technology into the internet. Proceedings of ACM HotNets-I Workshop, Princeton, New Jersey, USA, October 2002.
[19] A. Paepcke, S.B. Cousins, H. Garcia-Molina, S.W. Hassan, S.K. Ketchpel, M. Roscheisen, T. Winograd, Using distributed objects for digital library interoperability, Available from <http://dbpubs.stanford.edu:8090/pub/1998-57> 1998.
[20] J.R. Smith, S.F. Change, VisualSEEK, a fully automated content-based image query system, International Multimedia Conference, in: Proceedings of the fourth ACM International Conference on Multimedia, Boston, Massachusetts, 1997, pp. 87–98.
[21] R. Wilensky, Digital library resources as a basis for collaborative work, Journal of the American Society for Information Science 51 (2000) 228–245.
[22] R. Larson, C. Carson, Information access for a digital library: Cheshire II and the Berkeley environmental digital library, in: Proceedings of the 62nd ASIS Annual Meeting, November 1999.
[23] J. Vaidya, C. Clifton, Privacy-preserving k-means clustering over vertically partitioned data, SIGKDD 2003.
[24] A. Evfimievski, J. Gehrke, R. Srikant, Limiting privacy breaches in privacy preserving data mining, PODS 2003.
[25] R. Agrawal, Privacy in data systems, PODS 2003.
[26] I. Dinur, K. Nissim, Revealing information while preserving privacy, PODS 2003.
[27] W. Du, Z. Zhan, Using randomized techniques for privacy-preserving data mining, SIGKDD 2003.
[28] B.Y. Zhao, L. Huang, J. Stribling, S.C. Rhea, A.D. Joseph, J.D. Kubiatowicz, Tapestry: a resilient global-scale overlay for service deployment, IEEE Journal on Selected Areas in Communications (2003).
[29] S. Rhea, P. Eaton, D. Geels, H. Weatherspoon, B. Zhao, J. Subiatowicz, Pond: the OceanStore Prototype, in: Proceedings of the 2nd USENIX Conference on File and Storage Technologies (FAST '03), March 2003.

Ruzena Bajcsy is a professor in the Electrical Engineering and Computer Science Department at the University of California Berkeley. She was appointed Director of CITRIS (Center for Information Technology Research in the Interest of Society) at UC Berkeley on November 1, 2001. Prior to coming to Berkeley, she was Assistant Director of the Computer Information Science and Engineering Directorate (CISE), with a $500 million annual budget, at the NSF. Dr. Bajcsy is a pioneering researcher in machine perception, robotics and artificial intelligence. She is also a member of the Neuroscience Institute and the School of Medicine at the University of Pennsylvania. She is the former Director of the University of Pennsylvania's General Robotics Automation Sensing Perception Laboratory, which she founded in 1978. She has done seminal research in the areas of human-centered computer control, cognitive science, robotics, computerized radiological/medical image processing and artificial vision. She is highly regarded, not only for her significant research contributions, but also for her leadership in the creation of a world-class robotics laboratory, recognized worldwide as a premiere research center. She is a member of the National Academy of Engineering, as well as the Institute of Medicine. She is especially known for her wide-ranging, broad outlook in the field and her cross-disciplinary talent and leadership in successfully bridging such diverse areas as robotics and artificial intelligence, engineering and cognitive science. Dr. Bajcsy received her Master's and Ph.D. degrees in electrical engineering from Slovak Technical University in 1957 and 1967, respectively. She received a Ph.D. in Computer Science in 1972 from Stanford University. She then both taught and did research at U Penn's Department of Computer and Information Science where she began as an assistant professor and within 13 years became chair of that department. Prior to her work there, she taught as an instructor and assistant professor in the Departments of Mathematics and Computer Sciences at Slovak Technical University in Bratislava. She served as advisor to more than fifty Ph.D. recipients. In 2001 she received an honorary doctorate from the University of Ljubljana in Slovenia. In 2001 she was a recipient of the ACM A. Newell award. Discover Magazine named her to its list of the 50 Most Important Women in Science in November 2002. In April 2003 she received the CRA Distinguished Service Award and in May 2003 she become a member of the President's Information Technology Advisory Committee (PITAC).

Rick McGeer earned his Ph.D. in Computer Science from UC-Berkeley in 1989. From 1989 to 1991 he was a professor in the Computer Science Department at the University of British Columbia. In 1991 he returned to UC-Berkeley as a Research Engineer in the EECS Department. In 1993, together with Alex Saldanha, Luciano Lavagno, Alberto Sangiovanni-Vicentelli and Patrick Scaglia, he founded the Cadence Berkeley Labs where he served as a Research Scientist until 1999. In 1998, together with Alex Saldanha, he founded Softface, Inc., a successful software startup currently based in Walnut Creek. In February, 2003, entranced by the exciting possibilities of Cal's partnership with HP in the CITRIS program, Rick leapt at the chance to be part of two great research institutions and joined HP Labs as the CITRIS Scientific Liaison. Rick is the author of over 50 refereed technical publications, "Integrating Functional and Temporal Domains in Logic Design" (Kluwer, 1991), holds six patents and has won best paper awards at the International Conference on VLSI, the Cadence Technical Conference and the Hawaii International Conference on the System Sciences. He has served on numerous conference technical committees, and has served as General Chair of the IEEE Workshop on VLSI, the ACM/IEEE Workshop on Logic Synthesis, and the founding General Chair of the Tau Workshop series.