

Culture and Cyberinfrastructure: the need for a cultural informatics

Gregory Crane, Tufts University, Gregory.Crane@Tufts.edu

What are we trying to do? – the Challenge of Cultural Communication

I would like to focus on the problem of cultural communication: understanding cultures, our own and those of others very different from us, and communicating our understanding both of ourselves and of others to the rest of the world.

The importance of such cultural communication cannot be overstated. The war on terror is a war about culture — while material causes such as oil or land may contribute to the problems that we face, the acts of violence that we confront are only the most acute (and overtly destructive) symptoms of a broader conflict between cultures in a changing world. Such cultural conflicts extend well beyond the West and Islam (e.g., ethnic terrorism in Rwanda and Congo) and have been, in one form or another, endemic to human history, but the threats to U. S. security have clearly escalated. And if the war on terror has influenced our view of cultures, we must remember that culture stands at the core of human existence. Preserving and disseminating cultural heritage is a core responsibility for any government.

The challenges posed by terrorism have risen high on the U. S. agenda. Programs such as DARPA's Translingual Information Detection, Extraction and Summarization (TIDES) have acquired a new urgency, as intelligence agencies attempt to extract useful information from vast bodies of data. At the same time, the administration's proposed "We the people initiative" (<http://www.wethepeople.gov/>) — a \$100 million three year NEH effort — addresses the problem of articulating American history and culture. On its homepage Bruce Cole, NEH Chairman, explicitly cites the attacks on September 11 as a motivating cause for this effort. Many U. S. government institutions, including the Library of Congress, the Institute for Museum and Library Services, the National Endowment for the Arts, the Smithsonian Institution, the National Park Service and others, constantly labor to improve and disseminate knowledge about American history and culture.

Nevertheless, the rise of information technology has not only opened up new ways of analyzing cultural heritage materials but has created the need for a new area of study. Possible labels for this new field include "computing and the humanities" and "computational humanities" but I would suggest "cultural informatics" or some other term that emphasizes culture in the broadest possible sense, including not only the humanities and cultural heritage materials but regarding all forms of current culture as well.

Consider two broad applications:

Cultural awareness: How do we track changes in cultural values and attitudes not only in the past but in real time? Such a task is immense, for it assumes systems that can manage hundreds, if not thousands, of languages and cultural groups. We need to stress the importance of culture and language alike — information extraction may track telephone numbers or military appointments (e.g., General X is now in command of facility Y) but in this case, multiple languages describe trans-cultural objects: the information is the same and the language serves as a semi-transparent delivery medium. But the values which motivate humans most strongly – virtue, justice, even love and friendship – are not only fluid and resistant to reductive definition in any one culture, but vary substantially from one culture to another. Most inhabitants of the Western world shook their head when Osama Bin Laden cited a prophetic dream during a recorded video. Those of us who have worked with ancient Near Eastern or even classical Greek literature, however, realized that Bin Laden was following ancient habits of thought. We need instruments to assess culturally specific attitudes that are foreign to the discourse of international technocratic elites. We need to be able to assess how typical a particular sentiment may be (e.g., how commonly do conservative Saudis use dreams in this way? Is Bin Laden typical? Or is he self-consciously striking an archaic pose?)

Since small numbers of determined individual can potentially wreak immense havoc, we cannot predict where the next threat will originate. Security concerns thus augment broader interests in culture, demanding systems that can automatically analyse cultural materials in radically new ways.

Consider as one particular application of general semantic analysis a visualization system that tracked intensity of feelings about a particular topic across the globe. The system would take a range of textual materials from many different sources and languages and produce a measure of the feelings on a given topic. The system might use temperature gradients as a visualization metaphor, allowing viewers to track hostility towards the United States as it ebbs and flows across the world. Such a system could provide data for strategic planning, allowing the United States to identify problem areas as they emerge and before they generate violent attacks.

Such semantic analysis may be feasible for a constrained set of genres and languages, but a truly scalable system would not only have to manage a dizzying variety of languages but would also have to assess the relative rhetorical intensity of a given source: the same terminology can have radically different significance in a restrained bureaucratic document and in a hyperbolic partisan news story.

Spatial customization: Computing may become ubiquitous and the capitalism that shapes all of our lives pushes us to transcend our local surroundings, but many, if not most, human beings treasure a strong sense of place. The perceived rootlessness of capitalist and especially American life is arguably one of the most powerful concepts behind opposition to globalization in general and the United States in particular.

Connecting history to place has, historically, been a difficult task. Those who live in long established communities may have only the dimmest understanding about the generations who lived, struggled, suffered, laughed, walked and died in the areas through which they now walk each day. We can, however, imagine systems that would deliver

customized information to individuals as they move through a particular space. A system may alert the visitor that an object of interest (a Greek revival house built in 1847, a statue commemorating a hero of the anti-slavery movement, the home of a jazz musician) is only a few steps away, provide a survey of the names and occupations of those who lived in a particular building, historic images of their current location or dynamic VRML representations of the location in a particular period.

Such dynamic spatially customized data clearly has broad military and commercial applications, but military and commercial needs, however acute or potentially lucrative, are subsets of cultural analysis. Note that, as with purely textual information extraction, some crucial questions may be relatively easy to identify (e.g., location of possible ambush points, identification of customers for a particular product) but long term success may depend upon access to much more diffuse information about local practice and knowledge.

How do we accomplish cultural communication today?

At present, thousands of researchers in a variety of disciplines study cultures past and present. The vast majority of these researchers distribute the results of their work in diffuse papers (such as this position paper) and monographs. While almost all make use of electronic text processing and search various resources, most content themselves with basic searches of fragmented data sources.

What is new and why do we think it will be successful?

Such venues as the Text Retrieval Evaluation Conference, the Document Understanding Conference, Automatic Context Extraction, Cross Language Evaluation Forum, etc. are now documenting the capabilities of a range of emerging language technologies. If we could simply apply to conventional study the state of the art in technologies such as automatic multi- and single-document summarization, question answering, clustering, cross language information retrieval, named entity and relation detection, sense disambiguation, etc., we could revolutionize the ways in which we analyze culture.

Emerging language technologies have, however, been developed largely by engineers to work with modern genres (e.g., news publications produced and edited by professional journalists). We need to adapt existing techniques and develop new approaches to work with the diffuse genres from the cultures and languages of the world.

Assuming we are successful, what difference does it make?

Increased cross-cultural understanding/exchange can increase communication, reduce the impulse to violence and expand socio-economic connections. Thus, potential benefits include enhanced global peace and prosperity.

How long will it take and how will we know when we get there?

Cultural cyberinfrastructure requires two fundamental components.

First, we need knowledge resources. Some of these exist in digital form and can be federated into a useful resource network. Some digital sources will prove sufficiently inconsistent in form and/or quality that they need to be recreated. Most historical sources remain available only in print form. Many of the most useful potential sources are reference works that are expensive to enter (e.g., city directories) and have rich, but complex formats (e.g., dictionaries, grammars, encyclopedias). Automation can potentially reduce the costs of large-scale scanning to very low levels, but some tasks may require carefully entered and marked-up knowledge sources (c. \$500-\$1,000/mbyte). The DARPA TIDES program has studied the problem of boot-strapping new languages. Our initial estimate, based on our own experiences with a handful of languages, leads us to suggest that \$100 million would probably allow us to provide consistent initial coverage 100 major culture/language clusters (thus allowing us to consider Indian/English, UK/English, US/English). Actual costs would depend upon factors such as the complexity of the language, the availability of core resources (e.g., bi-lingual corpora, lexica, etc.) in print and/or electronic form, etc. The time required will depend upon the range of services developed: cross-language information retrieval and very rough machine translation may require relatively little time, but each language/culture cluster will pose its own problems. Nevertheless, three years should be enough for a competent team to provide basic language/culture services. Assuming two years of preparation, three cycles of grant competitions for language/culture teams, and the usual extensions/delays, we should be able to see major results in five years and a solid 100 language/culture infrastructure in 10.

Second, we need researchers who can assess the latest findings of various cultural disciplines and of emerging information technologies. We do not yet have the institutional framework to train the researchers with professional training in the analysis of culture and the possibilities of emerging technologies, language and otherwise. Cultural disciplines have not developed a strong cultural informatics – and these disciplines may never, as currently constituted, have the resources to support such a sub-discipline.

In an ideal scenario, departments in computer science, information science and various cultural domain disciplines would all train students and support faculty concentrating on various elements of cultural informatics. NSF support for cultural informatics should be aimed at helping this discipline develop the scientific and engineering rigor that it needs. While it will take five to ten years to develop a solid core of new researchers, the results of DLI-2 support suggest that we should begin to see research results within one to two years.