eCulture, Cyberinfrastructure, Virtual Research Environments and the Cultural Heritage of the Maghreb

Gregory Crane, Tufts University

The three neologisms in the title reflect three related phenomena. The term eCulture describes the nature of intellectual life in a world of high-speed, global neworks, intelligent services, and massive data – eCulture subsumes eScience, eResearch and other terms that describe formal academic work and includes phenomena such as wikis, blogs, video games. Virtual Research Environments (VREs) describe particular, generally academic configurations which aggregate data and services for particular communities – they can be seen as comprehensive digital libraries.[1] Cyberinfrastructure describes the general technological foundations on which VREs and eCulture are built.[2] If we were to compare the transformation of intellectual life to the creation of sedentary farming, eCulture would correspond to agrarian life, VREs to individual agricultural communities with their particular fields and crops, and cyberinfrastructure to the technologies of ploughing, seed production, individual farming and collective social structures for communal efforts.

eCulture, cultural heritage and language

Many of us have children who have grown up immersed in the 3d spaces – often inspired, for better or worse, by historical periods and foreign cultures. Virtual tourism in the past – especially well-documented pasts -- is well within our grasp and has arguably already begun. We have mature and improving methods with which to represent complex objects and spaces separated from us by time or space.

Language remains the great barrier to cultural understanding. Those who follow "religions of the book" find in words the meaning of life and their most profound guides for action. All of us fashion ourselves from the words of our cultures. In an emerging eCulture we not only have physical access to words from around the world and from every period of human history but we have new methods with which to grasp and explore their meaning. Machine translation, errorful as it may be, has begun to open new perspectives.[3] Google's Arabic to English translation is far from fluent but I can see on the publisher www.alwaraq.net that the *Thousand and One Nights* is the most commented on book and that there is audiovisual material about "Shakespeare's artistic skills." But even as such machine translation improves, it will only provide more generally a service that human translators have provided since writing was invented and the first Akkadian translations of Sumerian appeared four thousand years ago.

---

[1] An extensive discussion of VRE's can be found in (Fraser 2005).
[2] For more on the issue of the need for cyberinfrastructure beyond the sciences, please see (ACLS 2006).
[3] For a recent overview of the potential of machine translation, please see (Smith 2006).

In the *Faydrus* of Aflatun (Plato's *Phaedrus*), a speaker vehemently condemns written language as inert and as lifeless as a statue: we may ask whatever questions we please, but the written words will remain silent as stone. In eCulture, however, our books not only accompany us at all times but they know who we are and adapt themselves to our backgrounds and our needs. Our books can know what languages we have learned. They can know what we have read and what words we have seen and when and where we have seen them. They can remember what people, places, organizations and other background information we have encountered. They can prioritize what we do not know, infer from the questions that we pose what other questions may interest us. They can inform us, whether a particular reference to a person, place or topic puzzles us or we are casting a quick glance at the Arabic language newspapers of the day.[4] The task of learning Arabic or any other language will always be arduous but our books can be our guides, coaching us, reminding us of what we have learned, scanning for problems in our work, suggesting points on which we should focus our efforts. Our books become our helpers – beneficent *jinn*, perhaps not so clever as those which served Aladdin, but not so dangerous either.

One of Plato's major points in the *Phaedrus* is that disembodied ideas – for him the written word but for us more generally information – have no value or utility when stored in a physical medium. Ideas only matter insofar as they find their ways into the minds and the hearts of human beings. Customization allows users to describe the materials that wish to see. Personalization allows books to organize themselves to support the needs of their readers.[5] We may be accustomed to seeing these functions in e-commerce ("people who bought book X also bought books Y and Z") but there is nothing inherently commercial in these processes. We may as well have a system that tells us "people who looked up words A, B, and C, also had questions about words X, Y, and Z."

In a print library, a book in Arabic or Classical Greek is, at best, an object of respectful curiosity to the average speaker of English or Chinese. In a true digital library, the Arabic book translates itself, summarizes its contents, plots maps and timelines of the places and dates it contains, tells the stories of its main characters, and provides for human readers the network of background information that they need. And it helps them learn for as briefly, or as long, as they choose.

Cyberinfrastructure for Cultural Heritage Materials

Our cyberinfrastructure needs three core processes.

- Analog to text: While automatic speech recognition is crucial, we have much work to do on optical character recognition if we are to provide acceptable results for many cultural heritage materials. Even when our texts are cleanly printed in Roman characters, we need systems that know about historical languages and do not convert Latin words such as *tum* ("then") into similar English (e.g., *turn*). We

[4] For further exploration of these issues please see (Crane et. al. 2006a)
[5] Examples of personalization and customization found in the Perseus Digital library have been described in (Crane et. al. 2006b).

have made excellent progress with OCR of classical Greek but Arabic print needs considerable research.[6] In Arabic, in particular, we must deal with an even larger body of material still in manuscript form and not yet available in a modern printed edition. This process should extract as much information implicit in the page layout as possible – e.g., identification of headers, page numbers, footnotes, bibliographic entries, block quotes.

- Text to data: At the foundation are atomic processes, which classify words and phrases: named entity identification systems identify encyclopedic data such as people, places, organizations, etc. (is Washington a person or a place? If a place, which Washington is it?); linguistic analyzers provide part of speech or more elaborate morphological information. Information extraction systems then look for propositional data (e.g., Person-X at Place-Y). Syntactic and semantic analyzers combine named entity and morphological data to detect higher level structures (e.g., noun X is the object of verb Y; verb A has subjects who are human beings). This stage provides not only raw material for text mining but also the foundation for links between subject matter and background materials (e.g., links between a reference to Ahmed to an article about the particular Ahmed in question, links form an inflected Arabic verb to its morphological analysis and a dictionary entry).[7]

- One language to another: This includes not only machine translation but cross language information retrieval and advanced reading support for students of a language, including dictionary lookups, idiom identification, etc.[8] We also need tools that can automatically locate translations of a given text already available on-line and then align the translations to the source text as closely as possible.[9] This serves two audiences. Human readers find parallel source texts and translations useful – especially when there are links between many of the corresponding words and phrases. At the same time, parallel corpora provide fundamental material on which technologies such as machine translation draw.

The services above depend upon well-structured digital objects. Raw OCR output is not enough. Even cleanly produced PDF files are inadequate—they add some functionality (e.g., searching, ability to resize and expand) but they are digital incunabula, which mimic print and are not yet truly digital. The following features characterize post-incunabular objects that are designed as components of a digital library.[10]

- Content and display are separated: The separation of content from display is now a venerable architectural principle but remains praised in the breech rather than practiced. We need XML documents on which multiple services can build added

[6] There is a growing body of literature on the issues of OCR and document recognition with historical languages, please see (Leydier 2005) and for particular issues with Arabic, (Shahab, et. al. 2006)
[7] Preliminary research on named entity recognition in Arabic has been reported by (Nezda et. al 2006)
[8] For a review of some of the current challenges in CLIR, please see (Gey 2005).
[9] Some interesting research in this area has been reported by (Pouliquen 2003).
[10] For more on this topic, please see (Crane, et. al. 2006a)

value services and which are designed to be fully customized/personalized. The separation of content from display is essential if we are to create large collections that are deeply interoperable.[11]

- Documents are dynamic: Reference works, for example, are constantly updated. Updates can be manual (as in Wikipedia) or reflect automated processes (e.g., the population figures for a given city can be linked to an external database). In a scholarly environment, however, documents must also be versioned and scholarly citations must have time-stamps so that we can always reconstruct the evidence on which a given statement was originally based.

- Documents are recombinant: They are designed to disassembled into well-defined chunks, then recombined dynamically to create customized/personalized displays: thus, if an American reader calls up an Arabic section of Ibn Khaldun's *Muqaddimah*, the system might assemble, on the fly, multiple English translations, a list of the most important entries from a dictionary, key articles from several encyclopedias, etc.[12]

- Documents talk to each other: Marvin Minsky suggested that the day would come when no one would believe that the books in the library once did not talk to one another. Minsky may have imagined advanced artificial intelligence, but very useful conversations are already taking place between books as they decide how best to serve their human readers. Thus, when someone reading a Greek author in the Perseus Digital Library looks up a particular word, the Greek text tells the dictionary what passage, work and author is being read. The Greek lexicon can then check to see if it has a meaning for this word in this particular context and present that directly. If not, then the dictionary can highlight those senses which are attested for this particular work and/or author.

A Virtual Research Environment for Arabic

Between the idealized practice of eCulture and the generalized forms of cyberinfrastructure stand virtual research environments which, though virtual in name, actually deliver content and services to people in the real world. Tufts University has received support from the Department of Education to begin work on a virtual research environment to support those reading about Arabic and Islamic culture in English and working with the Arabic language directly. This work builds on infrastructure already developed for classical Greek and Latin and for the National Science Digital Library.

There are several components to this architecture.

- User models: These include the vocabulary which Arabic students have encountered and the more general knowledge base of people, places, organizations, and other term based topics that students, whether working Arabic

[11] For more on the importance of interoperability, please see (Van de Sompel, et. al, 2006)
[12] Please see (Kelly 2006) for some interesting possibilities in this area.

or English, have encountered.[13]

- Knowledge bases: Traditional reference materials – lexica, encyclopedias, grammars, indices, gazetteers – are knowledge bases insofar as their contents are machine actionable. We have locally two Arabic lexica entered so far (Salmoné and a two-volume edition of Lane), each of which needs added structure so that we can automatically generate links between inflected and the appropriate dictionary sections.[14] For English language background information, we will evaluate Wikipedia (which can be freely downloaded and analyzed).

- Information extraction: This includes term based services:

    - morphological analysis (e.g., parsing inflected forms and identifying dictionary entries from which they could be derived), for which we are using the Buckwalter Arabic Morphological Analyzer.[15]

    - named entity identification (e.g., determining when Lebanon refers to the country and when it describes one of the numerous towns named Lebanon in the US and elsewhere in the world). We developed an experimental system of our own that produces very good results[16] but plan to shift this function to a more general open source solution such as Sheffield's GATE (Generalized Architecture for Text Engineering[17]) or IBM'S UIMA (Unstructured Information Management Architecture[18]).

    - gazetteer lookup: Many terms – especially multi-word phrases – are sufficiently distinctive that we can provide good results simply by string lookups without complex disambiguation strategies. We have a service (SCALE) developed for the National Science Digital Library to identify technical terms that can efficiently identify multi-word terms in text (e.g., matching any instance of Ibn Khaldun, A Thousand and One Nights, etc.).[19]

    - machine translation: Google has released BETA versions of Arabic/English and English/Arabic machine translation. How useful is such a system in 2007?

- Institutional repository: We have chosen Fedora as the repository system in which we will deposit our digital objects. We chose Fedora because its model

---

[13] User modeling is an area of extensive research, for one recent application, please see (Brusilovsky 2005).
[14] A growing number of projects are exploring the creating of Arabic language resources, including the NEMLAR project, please see http://www.nemlar.org.
[15] (Buckwalter 2004).
[16] For a discussion of the experimental system developed for Perseus, please see (Crane and Jones 2006).
[17] http://www.gate.ac.uk/.
[18] http://www.uima.info/; http://sourceforge.net/projects/uima-framework.
[19] http://dca.tufts.edu/scale/

also allows us to integrate the high level services listed above on which readers will depend.[20] We are closely following the Mellon-funded Object Re-use and Exchange Project, which is developing methods to combine objects from multiple repositories to create new aggregations of information such as were described in the section on cyberinfrastructure.[21]

- Grid services: As academic disciplines begin to merge their data into large, international collections, grid technologies have emerged as a potential foundation for (1) long-term preservation[22], (2) access to high performance computation for demanding tasks such as text mining[23], (3) creation of shared repositories within individual grid environments, (4) integration of multiple repositories from grid systems around the world. The Fedora repository already has an interface whereby we can store objects directly in the Storage Resource Broker distributed storage environment developed by the San Diego Supercomputer Center. Discussions are on-going to establish mechanisms whereby humanities collections from the US, the UK and Germany are automatically replicated, with each Grid home to its own collections and with up-to-date copies of collections from other grids.[24]

- Evaluation: Our initial evaluation in fall 2007 will focus on intermediate and advanced classes in Arabic and classes on Arabic literature and culture taught in English at Tufts University. We are, however, looking to begin discussions about broader evaluations and collaborations.

- Collections: Enough materials are freely available on-line with which to begin research, development and evaluation of many (though not all) functions. The Perseus Digital Library has over twenty years developed collections on topics including Greco-Roman culture, early modern Europe and 19th century British and American history and literature. We are eager to contribute in whatever way we can to a digital library on the Maghreb Digital Library.

Collection development

What would the cultural heritage component of a Maghreb Digital Library contain? Where are the particular cost/benefit tradeoffs for the materials relevant to this collections? Searchable texts will, for example, probably be more expensive to produce than 19th century English – or even classical Greek. On the other hand, the historical and literary heritage of the Maghreb may deliver even greater benefits to the nations of North

---

[20] For more on Fedora, please see (Lagoze, et. al. 2006)
[21] http://www.openarchives.org/ore/.
[22] For more on the grid potential for data preservation, please see (Moore 2006)
[23] Please see, for example, (Watry, et. al. 2006)
[24] For a detailed description of this work within Germany, please see http://www.textgrid.de/index.php?id=konferenzen&L=5

Africa than vast, open access libraries of British and American materials confer upon the United Kingdom and the United States.

Some collection development suggestions:

- Develop a general library covering every aspect of Maghreb culture from antiquity to the present. This should include a prioritized library of texts, as many reference materials as are available, images documenting objects and places, both historical and new, geospatial data and especially historical gazetteers, multimedia representations of poetry, music and other cultural productions, etc.

- Pick several areas on which to provide deep coverage in order to understand the challenges that will emerge as the collection grows over time. These should include geo-spatial topics (e.g., one or more cities, smaller sites and individual buildings), people (e.g., as many sources as possible to illustrate the lives of several key figures), genres (e.g., one or more forms of literature, art, music, etc.)

- Establish a core of open access and, where possible, open source content to encourage scholarly analysis and innovative derivative works. Classical studies has fostered a small, but active core of faculty with a sustained record of developing digital infrastructure. This community has found an open access at minimum and open source at best policy essential for progress. Scholars working on the Maghreb in a digital environment will probably, as they emerge, come to similar conclusions and the Maghreb Digital Library should be designed to accommodate the decentralized, rapid nature of research in a global digital environment.

References

American Council of Learned Societies Commission for Cyberinfrastructure for the Humanities and Social Sciences. (2006). “Our Cultural Commonwealth: The final report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences."
http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

Brusilovsky, P, S. Sosnovksy, and O. Shcherbinina. (2005). "User modeling in a distributed E-learning architecture." *User Modeling 2005*, pp. 387-391.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer Version 2.0. LDC2004L02.
(http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004L02)

Crane, Gregory, D. Bamman, and A. Babeu. (2006a) “ePhilology: When the Books Talk to Their Readers.”, forthcoming in *Blackwell Companion to Digital Literary Studies*, , New York London : Basil Blackwell.
http://dl.tufts.edu//view_pdf.jsp?urn=tufts:facpubs:gcrane-2006.00003

Crane, Gregory, et. al. (2006b). "Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries." *ECDL 2006,* pp. 353-66. http://dl.tufts.edu//view_pdf.jsp?urn=tufts:facpubs:gcrane-2006.00002

Crane, Gregory and Alison Jones. "The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection." *JCDL 2006,* pp. 31-40.

Fraser, Michael A. (2005). "Virtual Research Environments: Overview and Activity. *Ariadne*, Issue 44, http://www.ariadne.ac.uk/issue44/fraser/.

Gey, Fredric C. and Noriko Kando and Carol Peters. (2005). "Cross Language Information Retrieval: the Way Ahead." *Information Processing & Management*, May, 41 (3), pp. 415-31.

Kelly. K. (2006). "Scan This Book!" *New York Times Magazine.* http://www.nytimes.com/2006/05/14/14publishing.html

Lagoze, C., S. Payette, E. Shin, and C. Wilper. (2006). "Fedora: architecture for complex objects and their relationships." *Int. J. Digit. Libr.*, 6(2):124–138.

Leydier, Yann, F. LeBourgeois, and H Emptoz. (2005). "Textual Indexation of Ancient Documents." *DocEng 2005*, pp. 111-117.

Moore, Reagan. (2006). "Building Preservation Environments with Data Grid Technology." *American Archivist*, 69 (1), pp. 139-158.

Nezda, L., et. al. (2006). "What in the World is a Shahab? Wide Coverage of Named Entity Recognition for Arabic." *Proceedings of the 2006 Language Resources and Evaluation Conference (LREC 2006).* Genoa, Italy.

Pouliquen, Bruno, R. Steinberger, and C. Ignat. (2003). "Automatic Identification of Document Translations in Large Multilingual Document Collections." *Proceedings of the International Conference 'Recent Advances in Natural Language Processing' (RANLP'2003)*, pp. 401-408. http://arxiv.org/pdf/cs.CL/0609060

Shahab, S.A., W. G. Al-Khatib, and S.A. Mahmoud. "Computer Aided Indexing of Historical Manuscripts." *CGIV 2006*, pp. 287-95.

Smith, David. (2006). "Debabelizing Libraries: Machine Translation by and for Digital Collections." *D-Lib Magazine*, March, 12 (3), http://www.dlib.org/dlib/march06/smith/03smith.html

Van de Sompel, Herbert, et. al. (2006). "An Interoperable Fabric for Scholarly Value Chains." D-Lib Magazine, 12 (10), http://www.dlib.org/dlib/october06/vandesompel/10vandesompel.html

Watry, P., R. R. Larson and R. Sanderson. (2006). “Knowledge Generation from Digital Libraries and Persistent Archives.” ECDL 2006, pp. 504-7.