A Revolutionary Approach to Humanities Computing?: Tools Development and the *D2K* Data-Mining Framework

J. Stephen Downie *(jdownie@uiuc.edu)*
University of Illinois at Urbana-Champaign
John Unsworth *(unsworth@uiuc.edu)*
University of Illinois at Urbana-Champaign
Bei Yu *(beiyu@uiuc.edu)*
University of Illinois at Urbana-Champaign
David Tcheng *(dtcheng@ncsa.uiuc.edu)*
University of Illinois at Urbana-Champaign
Geoffrey Rockwell *(georock@mcmaster.ca)*
McMaster University
Stephen J. Ramsay *(sramsay@uga.edu)*
University of Georgia

Introduction

A new set of humanities computing (HC) research projects are underway that could revolutionize how the HC community works together to build, use, and share HC tools. The set of projects under consideration all play a role in the development work currently being done to extend the *D2K* (Data-to-Knowledge)[1] data-mining framework into the realm of HC. **John Unsworth** and **Stephen J. Ramsay** were recently awarded a significant Andrew W. Mellon Foundation grant[2] to develop a suite of HC data-mining tools using *D2K* and its child framework, *T2K* (Text-to-Knowledge). Drs. Unsworth and Ramsay, along with research assistant, **Bei Yu**, are working closely with **Geoffrey Rockwell**. Dr. Rockwell is the project leader for the *CFI* (*Canada Foundation for Innovation*) funded project, *TAPoR* (*Text Analysis Portal for Research*)[3], which is developing a text tool portal for researchers who work with electronic texts. **J. Stephen Downie** and **David Tcheng**, through their work in creating the *International Music Information Retrieval Systems Evaluation Laboratory* (*IMIRSEL*)[4], are leading an international researchers group to develop another *D2K* child system called *M2K* ("Music-to-Knowledge"). This panel session demonstrates how all of these projects come together to form a comprehensive whole. The session has four major themes designed, through presentations and demonstrations, to highlight individual the project components being developed and their collective impact on the future of HC research. These themes are:

1. *D2K* as the overarching framework
2. *T2K* and its ties to traditional text-based HC techniques
3. *M2K* and its ties to multi-media-based HC techniques
4. The issues surrounding the HC community's development, validation, distribution, and re-use of *D2K*/*T2K*/*M2K* modules.

Participants

J. Stephen Downie, Graduate School of Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC)
John Unsworth, GSLIS, UIUC
Bei Yu, GSLIS, UIUC
David Tcheng, National Center for Supercomputng Applications (NCSA), UIUC
Geoffrey Rockwell, School of the Arts, McMaster University
Stephen J. Ramsay, Department of English, University of Georgia

Presentations, Demonstrations, and Discussions (in order)

Overview of the NORA (No One Remembers Acronyms) project

John Unsworth

For decades, humanities computing researchers have been developing software tools and statistical techniques for text analysis, but those same researchers have not succeeded in producing tools of interest to the majority of humanities researchers, nor (with the exception of some very recent work in the Canadian *TAPoR* project) have they produced tools that work over the web. Meanwhile, large collections of web-accessible structured texts in the humanities have been created and collected by libraries over the last fifteen years. During that same time period, with improvements database and other information technologies, data-mining has become a practical tool, albeit one mostly used in business applications. We believe data-mining (or more specifically, text-mining) techniques can be applied to digital library collections to discover unanticipated patterns, for further exploration either through traditional criticism or through web-based text analysis.

Existing humanities e-text collections from Virginia, Michigan, Indiana, North Carolina, and other research universities form the corpus for the project. *NORA* brings NCSA's *D2K* data-mining architecture to bear on the challenges of text-mining in digital libraries, with special emphasis on leveraging markup, and on visualizations as interface and as part of an iterative process of exploration.

Introduction to the D2K framework

David Tcheng

Released in 1999, *D2K* was developed by the Automated Learning Group (ALG) at NCSA. *D2K* has been used to solve many problems for both industry (e.g., Sears, Caterpillar, etc.) and government agencies (e.g., NSF, NASA, NIH, etc.). Academic uses include bioinformatics, seismology, hydrology, and astronomy. *D2K* uses a data flow paradigm where a 'program' is a network (directed graph) of processing modules. Modules can be 'primitive', defined as a single piece of source code that implements a single well defined task, or can be 'nested' meaning it is defined as a network of previously defined *D2K* modules. Decomposition of programs into modules that implement a well defined input-output relationship promotes the creation of reusable code. Nesting modules into higher-level modules helps to manage complexity. *D2K* parallelizes across any number different computers by simply running a copy of "D2K Server" on each available machine. The *D2K* software distribution comes as a basic *D2K* package, with core modules capable of doing general purpose data-mining, as well as such task-specific add-on packages as text analysis (T2K), image analysis (I2K), and now music analysis (M2K).

Introduction to T2K

Bei Yu

Similar to many data-mining tools, *T2K* has implemented a number of automatic classification and clustering algorithms. Compared to the commercial text mining tools, for example SAS Text Miner, *T2K* has richer NLP preprocessing tools, especially after its integration with GATE. Tools include: stemmer, tokenizer, PoS-tagger, data cleaning and named-entity extraction tools. The clustering visualization is tailored for thematic analysis. On one hand, *T2K* provides a text mining platform for the HC community. On the other hand, *T2K* is also a platform to automate the HC research results and thus facilitate their applications to the text mining community in general. For example, most of the text mining tasks are still topicality oriented, but the affect analysis has emerged in the last couple of years. The affect of a document includes the subjectivity/objectivity, the positive/neutral/negative attitude, and the strength of emotions, etc. Some researchers have adapted stylistic analysis techniques from HC to analyze customer reviews. The found non-thematic features can also be used as predictors for document genre, readability, clarity and many other document properties.

The TAPoR portal and D2K

Geoffrey Rockwell

TAPoR has released an alpha of the portal and will have the beta ready by June 2005. The portal is designed to allow researchers to run tools (which can be local or remote web services) on texts (which can be local or remote.) The *TAPoR* portal has been designed to work with other systems like *D2K* in three ways:

1. Particular tools or chains of tools can be 'published' so that they are available as post-process tool right in the interface of another system. Thus one can have a button that appears on the appropriate results screens of a *D2K* process that allows the user to pass results to *TAPoR* tools.
2. The portal has been released as open source and we are working on models for projects to run customized versions of the portal that work within their environment.
3. The portal can initiate queries to remote systems and then pass results to other *TAPoR* tools. Thus users can see tools like *D2K* (where they have permission) within their portal account.

The Tamarind project and D2K

Stephen J. Ramsay

Tamarind began with the observation that the most basic text analysis procedure of all — search — does not typically operate on the text archive itself. It operates, rather, on a specially designed data structure (typically an inverted file or pat trie index) that contains string locations and byte offsets. Tamarind's primary goal is to facilitate access to analytical data gleaned from large-scale full text archives. Our working prototype of Tamarind, for example, can quickly generate a relational database of graph properties in a text which can in turn be mined for structural information about the texts in question. Tamarind creates a generalized database schema for holding text properties and allows you to specify this structure as one that should be isolated and loaded into the database. Work is proceeding on a module that will allow the user to load a Tamarind database with millions of word frequency data points drawn from several gigabytes of encoded data. Unlike existing tools, this newest module includes information about where those counts occur within the tag structure of the document (something that is impossible to do without the raw XML). For the purposes of this project, we intend to use *D2K* and *T2K* as the primary clients for Tamarind data stores.

The M2K project

J. Stephen Downie

M2K is being developed to provide the Music Information Retrieval (MIR) community with a mechanism to access a secure store of copyright-sensitive music materials in symbolic, audio and graphic formats. *M2K* is a set of open-source, music-specific, *D2K* modules jointly developed by members of the *IMIRSEL* project and the wider MIR community. *M2K* modules include such classic signal processing functions as Fast Fourier Transforms, Spectral Flux, etc. In combination with *D2K*'s built-in classification functions (e.g., Bayesian Networks, Decision Trees, etc.), the *M2K* modules allow MIR researchers to quickly construct and evaluate prototype MIR systems that perform such sophisticated tasks as genre recoginition, artist identification, audio transcription, score analysis, and similarity clustering.

John Unsworth and **J. Stephen Downie** will lead a wrap-up and future work open-forum discussion: For ambitious, multi-institutional projects like those presented in this panel many issues arise that can affect the sustainability and impact of the projects. In particular, the issues surrounding the HC community's development, validation, distribution, and re-use of *D2K*/*T2K*/*M2K* modules will be addressed.

---

1. See <http://alg.ncsa.uiuc.edu/do/tools/d2k>.
2. See <http://www.news.uiuc.edu/news/04/1025mellon.html>.
3. <http://www.tapor.ca/>
4. See <http://music-ir.org/evaluation>.