1.0) Statement of CDL significance and impact

Chinese writing has been in continual use for more than 3,000 years, and Chinese-related writing systems are in use today by perhaps one-fifth of the world's population. Chinese and Chinese-derived (CJKV = Chinese, Japanese, Korean, Vietnamese) scripts preserve knowledge from many times and many places, but these scripts are so complex that they present immense challenges to digitization efforts. Because CJKV characters have traditionally not been encoded at a distinctive-feature level (i.e., the *character* is not the fundamental element of the writing system, but rather, the *stroke*), CJKV character-sets are open-ended, and writers take it for granted that for some purposes, writing with a computer is necessarily imperfect. The present proposal seeks to promote a practical and fundamental solution to this long-standing problem, to provide online tools for the use of international standards bodies, and ultimately to put these powerful tools in the hands of all who wish to use them. This work leads the way toward more precise and useful digitization of CJKV humanities collections, and promotes the international cooperation necessary to make the knowledge in these vast libraries available worldwide.

The *Character Description Language* (CDL) specification defines an XML application for describing any CJKV script entity (Bishop & Cook, 2003: <http://www.wenlin.com/cdl/>). Given the size and complexity of the current (Unicode 5.0) CJKV character-set (a.k.a. *UniHan*) which encodes some 72,000 characters, the task of maintaining and augmenting the international encoding is growing more difficult all the time. The CDL specification submitted to US and international standards bodies (Unicode and ISO/IEC 10646/WG2/IRG) inspired a great deal of work geared toward adoption of the CDL methods. At present, the only complete implementation of CDL is the C programming language implementation by Wenlin Institute (the educational software company owned by co-author Bishop, inventor of CDL). *Wenlin* is well known in Chinese second-language education as the preëminent software application for learning and studying Chinese (ancient and modern). Wenlin Institute is the maintainer of the Univ. of Hawaii's electronic *ABC Chinese-English Comprehensive Dictionary* data (DeFrancis et al.). The CDL database itself has been developed over the last decade in conjunction with co-author Cook (post-doctoral research fellow in the UC Berkeley Linguistics Dept., International Computer Science Institute, and co-editor of the *Unicode Standard 5.0*). The authors of this proposal have collaborated in the creation of some 56,000 CDL descriptions, and approximately 16,000 descriptions remain to be created, to cover the current encoded character-set. The present proposal aims at opening up the CDL system more widely, so that members of international standards bodies can employ CDL in their work to encode new characters and to secure and refine the mappings of existing characters. In the proposed funding period the proposers will create a prototype of an online centralized repository (MySQL database) of the existing CDL XML data, and complete a prototype version of the Wenlin client application for interacting with the server-side database. This client will be provided to the members of the international standards bodies contributing to *UniHan* development, and they will receive training in its use and access to the online CDL data. Long-term goals of the CDL project include creation of a public CDL interface to the Unicode Consortium's public *UniHan* database, and opening up the CDL source code itself to collaborative development, in an open-source framework compatible with Wenlin Institute's educational software business model. The proposed project will help to raise the profile of CDL among institutions and businesses with a vested interest in promoting long-term computing stability, and it is anticipated that present and future Unicode Consortium members will provide support for future open-source development.

2.0) Table of contents

[List all parts of the application and corresponding page numbers.]

The Character Description Language (CDL) Digital Humanities Start-up

3.0) List of CDL Project participants

[On a separate page, list in alphabetical order, surnames first, all project participants and collaborators and their institutional affiliations, if any. The names on this list should match the names mentioned in the staff section of the project's narrative description. The list is used to ensure that prospective panelists and reviewers have no conflict of interest with the project that they will be evaluating. This list should include advisory board members, if any.]

Project Leaders	
Bishop, Thomas E.	Wenlin Institute, Inc. (CEO and lead programmer); ABC Dictionary Project, University of Hawaii
Cook, Richard S.	Post-doctoral research fellow, Dept. of Linguistics, University of California, Berkeley; Artificial Intelligence Group, International Computer Science Institute; Unicode Consortium
Advisory Board Members	
Anderson, Deborah	Researcher, University of California, Berkeley Dept. of Linguistics; Script Encoding Initiative
Jenkins, John	Technical Director, Unicode Consortium, Apple Computer, Inc.
Lu Chin	Hong Kong Polytechnic University; Rapporteur, ISO/IEC JTC1/SC2/WG2/IRG
Lunde, Ken	Adobe Systems, Inc.
McGowan, Richard	Vice President, Unicode Consortium; Script Encoding Initiative
Whistler, Kenneth	Technical Director, Unicode Consortium, Sybase, Inc.

4.0) CDL Project Narrative

[Applicants should provide an intellectual justification for the project and a work plan. Narrative descriptions are limited to fifteen double-spaced pages. All pages should have one-inch margins and the font size should be no smaller than eleven point. Use appendices to provide supplementary material such as detailed work plans and résumés for project participants. The narrative should address the long-term goals for the project as well as the start-up activities that the Digital Humanities Start-Up Grant would support. Applicants should keep in mind the criteria (listed below) used to evaluate proposals. Provide a detailed project description that addresses the following topics:]

4.1) Enhancing the humanities through the use of CDL technologies

[Provide a clear and concise explanation of the start-up activities and the ultimate project results noting their value to scholars, students, and general audiences in the humanities. Describe the scope of the project activities, the relationship of the project to other published and ongoing work in the field, and major issues to be addressed. Applicants should provide a rationale for the compatibility of their methodological approach with the intellectual goals of the project and the expectations of its users. NEH views the use of open source software as a key component in the broad distribution of exemplary digital scholarship in the humanities. If either the start-up project or the long-term project is not predicated on generally accessible open source software, explain why and also explain how the Endowment's dissemination goals will still be satisfied by the project.]

The CDL project is directed in particular at resolving long-standing problems with the digitization of the Chinese, Japanese, Korean, and Vietnamese scripts (CJKV, or often simply shortened to CJK). The proposed start-up project will promote the collaborative use of the innovative CDL font technology, for the building of international computing standards essential to the stable function of all modern software, and for the accurate digitization and preservation of CJK documents and libraries. The CDL project benefits all computer users with an interest in CJK texts, or with an interest in dealing with CJK partners, since CDL has core applications for information input, storage, and retrieval. CDL improves data-management practices in the development of international standards and in the usage of end-users. CDL ensures the integrity of those standards, enriches the possibilities for end-user content creation, and therefore brings new richness to online digital humanities resources.

The CDL specification (Bishop & Cook, 2003) defines a standard (and standards-based) method for describing script entities (letters, characters, punctuation marks, symbols, etc.). The CDL method was developed specifically for CJK, but is not simply applicable to CJK. CDL has utility in the management of data for any character set for which there is some underlying reuse of basic components. The CDL XML application for CJK is based on a simple set of traditional stroke types, positioned in a standard grid space. Combinations of strokes define higher-level units, which may be reused in other CDL descriptions. The set of stroke types is termed "the ABC's of CJK", in that these strokes are the basic elements of CJK writing, rather like the alphabet is in English writing.

The problem solved by CDL is best explained by analogy with English. Imagine that instead of typing the letters of the alphabet on your computer keyboard to write English words, you were instead required to have a keyboard with one key for each *word* in the English language. This would be a big English keyboard indeed, for several reasons. There are many more words in English than there are letters in the alphabet, and the notion of "English word" is not well-defined: there are many different varieties of English, and each variety may have unique words and unique spelling rules. As English lexicographers must grapple with the fuzzy definitions of "English" and "word" in the compilation of their dictionaries, they often limit themselves to a specific genre or specific genres of English written in a specific time and place with specific orthographic rules. The situation in CJK encoding is rather similar to this, in that the CJK *character* (comprised of strokes) is rather like the English *word* (comprised of letters). The CJK *character* is ill-defined for precisely the same reasons that the English *word* is ill-defined. If English writers were encumbered by the restrictions currently afflicting CJK writers, the situation would be obviously intolerable.

For example, some English writers would not be able to spell their surnames or personal names, and they would be unable to use Shakespearean spellings in production of editions of Shakespeare, unless those words or spellings happened to be in a dictionary, and that dictionary also happened to have been used as a source for a computer encoding (assigning a unique number to each word type). Spellings from the works of Shakespeare, being fairly well known, would surely be on everyone's computer; but quirky personal or place names, or words in use by lesser-known writers, would surely be lacking. For example, we would be unable to mention in the present Unicode 5.0 proposal document the fact that Shakespeare reportedly spelled his name variously in his own day as "Shakspere, Shaksper, Shaxper, and Shake-speare", without resorting to some non-standard text representation practice. The digitization of such documents would be hindered by the lack of prior lexicographic and encoding work, and this presents an unacceptable and unnecessary bottleneck. Given that a goal of Unicode is to provide an international CJK character encoding, transcending temporal and political boundaries, Unicode does not have the luxury of narrow scope, and in fact the problems faced in Unicode's CJK encoding are greatly compounded, requiring in effect innovative and comprehensive lexicographic work to be the gating constraint on the digitization of texts. Clearly, unless an adequate system such as CDL for computerized encoding and structured combination of the minimally distinctive features of the script is developed and promoted for wide adoption, the short-term progress and long-term success of CJK digitization projects must remain limited.

By design (and by *necessity*), Unicode unifies (i.e. lumps together) many variants of CJK characters that differ in small ways, such as the presence or absence of a dot, or whether stroke segments are joined or separated. For some purposes such variations may be very important. For example, different national standards sometimes assign different official stroke counts to the "same" character, and the stroke counts determine the organization of dictionaries, etc. For another example, historical Chinese texts can often be dated on the basis of a single stroke being intentionally omitted in taboo avoidance of the name of the reigning emperor. Unicode has recognized the need for identifying variants that share a code point, and for this purpose has established the "variant selector" mechanism. The need remains, however, to assign a precise meaning to each combination of code point and variant selector (see Hiura & Muller 2006: *UTS #37*, < http://www.unicode.org/reports/tr37/>). CDL is ideal for this purpose. Each code point can have multiple CDL descriptions, one description for each variant, uniquely identified by a selector.

Of course, there will always be relatively rare or obscure variants that have not (yet) been assigned standard variant selectors, as well as characters that have not (yet) been assigned Unicode code points. CDL can be used to represent such characters directly without the use of code points or variant selectors, to feed candidates into the encoding and variant-mapping processes. For example, a CDL description can be included directly in the text, which might employ a higher-level markup such as XHTML. If the displaying software is CDL-enabled, it can display the character simply by interpreting the CDL, without the need for a customized font, without the need for a *Private-Use Area* code point, and without the need for an embedded graphical image. When end-users are able to create CDL descriptions of characters, encoded or not, and can embed these in online documents, web spiders crawling the internet can collect them automatically, and use associated metadata to feed these descriptions into encoding and variant-mapping processes.

Another exciting (and still highly theoretical) application of CDL is to the problem of CJK *optical character recognition* (OCR). Current OCR technologies, when applied to CJK text, yield largely binary results: either a printed character is identified with an encoded character, or it is not. Where matching fails or is imperfect, either a com-

pletely wrong character is selected, or else OCR fails completely. Using CDL, the results of partial or failed OCR matches become meaningful. Just as the inability to recognize a single letter in OCR of English text might not result in failure to recognize the word, and just as the absence of an English word from a spell-checker's dictionary needn't signal complete OCR failure, so too OCR of unencoded or damaged CJK characters can succeed where current CJK OCR fails, by including CDL in the OCR output.

The applications of CDL technology for scripts beyond the CJK world are just as important. The CDL team has discussed this with researchers working on other scripts, contributing to the on-going development of Unicode to handle digital text representation of the scripts of the world. For example, there has been growing interest in developing CDL schemes for Cuneiform scripts, Tangut (西夏 Xīxià), Egyptian Hieroglyphs, Mayan, and other complex scripts with componential basis and special component and character positioning requirements. Such scripts are similarly limited by the open-endedness of their character sets due to historical and local variation, and due to the ill-defined nature of the bounds of the higher-level units of writing. There are, nevertheless, basic script elements identifiable in each of these scripts, that can be employed in conjunction with CDL to remove the encoding bottleneck, and to empower the paleographers, lexicographers and linguists who must ultimately seek to resolve these problems.

CDL is a core infrastructure technology, providing a rock-solid framework for data structure, data storage and data interchange, and CDL should be adopted internationally in work to digitize and preserve humanities collections and paper and archeological archives. Because CDL is *pure Unicode* and *pure XML*, and because its applications for CJK are clearly based on *traditional orthographic standards* active across CJK scripts (as evident in the finite set of stroke types appearing in the representative glyphs in Unicode code charts), *the CDL standard is completely standards-based*, and poised to have a significant impact in the international community. This contrasts with existing "analogue" font technologies (TrueType, PostScript, OpenType), which make use of machine-focused technologies, the elements of which, although effective at producing the character shapes on various output devices, sever the connection to the actual human orthographic practices underlying those shapes. A conventional TrueType font may represent the same characters as a given CDL font, and that TrueType font may support a certain limited reuse of components. But the TrueType font does not preserve information about the many levels of componential structure of the characters, nor does it know anything about traditional stroke types or stroke order. Critical indexing information is built right into the heart of the CDL font, whereas in analogue fonts bulky external indices are simply tacked on later by an input method creator in a process separate from the typography. Thus, CDL technology bridges a significant gap, putting real intelligence into fonts, and giving CJK digitization projects the freedom which roman-based orthographies take for granted. The returns on small initial investments in this technology will be large indeed, and shared globally, as significant improvements are made to electronic data stability, electronic data content, and electronic data access.

4.2) History and duration of the CDL project

[Provide a concise history of the project, including information about preliminary research or planning, previous related work, previous financial support, publications produced, and resources or research facilities available. It is anticipated that work on projects initiated during the term of a Digital Humanities Start-Up Grant will continue after the period of the grant. The applicant should describe plans for that work and probable sources of support for subsequent phases of the project.]

The project proposers have been breaking old and new ground in Chinese computing for the last two decades, and have been collaborating since the early 1990's, to resolve many problems critical to Chinese information processing. Wenlin Institute, Inc. has been a leader in promoting the use of Unicode in CJK computing, and Wenlin was among the first developers to implement support for the vast Unicode 3.1 character-set (which added more than 40,000 new CJK characters to the already large repertory). The CDL system arose out of gradual refinements to Wenlin Institute's software for learning Chinese (in particular, out of software written to teach students the proper stroke order of handwritten characters), and the applications of this technology for resolving long-standing computing problems gradually became apparent.

Wenlin Institute, Inc. has been a California corporation since 1996, and has a well-established reputation in the educational software industry. Wenlin Institute, Inc. is itself not a "start-up" (which often connotes a new business in search of a working business model), since its revenues derive from sales of mature software. However, its CDL technology branch (with special emphasis on building international and institutional cooperation to solve a thorny shared computing problem in an open-source environment) certainly does fit this notion of "start-up". Wenlin's CDL technology is a powerful tool in search of a ecological niche in which it can flourish, and it would be wise for the international encoding community which so badly needs this system, to share some of the development burden. In fact, appreciation of the wider applications of the CDL system arose in collaboration, with contributions from end-users, and this is the best way for CDL to continue.

Beginning in 1998 the Wenlin Unicode text editor and the unpublished and undocumented CDL font editing system were used by Dr. Cook in preparation of large lexical databases for the NSF and NEH-funded STEDT Project at UC Berkeley. This database work resulted in publication of the first-ever digitization of important ancient Chinese lexical sources (Cook 2003), and resulted also in contributions of extremely large mapping tables (with nearly 100,000 records) to the Unicode Consortium's public *UniHan* database. The *UniHan* database (<http://www.unicode.org/Public/UNIDATA/Unihan.html>) at present contains more than one million records, and is maintained by Dr. Cook in collaboration with John Jenkins of Apple Computer, Inc. Existing *UniHan* data has undergone and is undergoing extensive proofing and continual validation using the CDL system. Wenlin Institute has donated its software for the use of Unicode editors and staff, and has also donated CDL data for use in publication of the *Unicode Standard*, including both Traditional-to-Simpified mapping tables, and also data employed in the sorting of the Radical/Stroke charts appearing in the new *Unicode Standard 5.0* book. Wenlin Institute has contributed to international standards work directly, sending Mr. Bishop to ISO/IEC 10646 meetings at its own expense, and has received support from Kyoto University (*Text Encoding Initiative,* Dr. Christian Wittern) to promote CDL technology in Japan.

The elements of the CDL system were first publicly specified in two core documents (Bishop & Cook, 2003; these documents are also included in the Appendix, §6.0):

- **"A Specification for CDL: Character Description Language"** .
  <http://www.wenlin.com/cdl/cdl_spec_2003_10_31.pdf >
- **"Character Description Language (CDL): The Set of Basic CJK Unified Stroke Types"**
  <http://www.wenlin.com/cdl/cdl_strokes_2004_05_23.pdf>

That initial documentation grew in stages to become an entire website promoting the use of CDL technology (<http://www.wenlin.com/cdl/>), and tracking CDL-related developments in the international encoding community. That website includes links to a number of encoding proposals and other documents submitted to Unicode and ISO/IEC 10646, introducing the power of CDL technology to the members of these standards organizations. CDL technology has resulted in significant refinements to *UniHan* mapping tables, in the correction of numerous errors in the international encoding standard, and in the addition of a new block of CJK Stroke characters (forthcoming, in Unicode 5.1). This latter work resulted from IRG work inspired by our initial specification of CDL, and the new block of stroke characters is intended to make the use of CDL more intuitive for end-users.

The present proposal seeks to take CDL work one step further, to open up the CDL system more widely as a suite of tools not only for the international encoding community, but also to make these tools available to collaborators and end-users worldwide. It is our firm belief that once the power of this technology is appreciated, it must naturally come to be used widely, and that wide adoption of the technology is the first obstacle to be overcome in seeking sustainable open-source development. This technology will attract new members to the Unicode Consortium, members who will see, as current members have begun to see, that application of CDL to CJK problem-solving should be a high priority, and is best undertaken in a collaborative open-source framework. NEH seed funding will help to get this project online, and help to make it accessible to the wider audience that will sustain it in the long run.

4.3) CDL Project Staff

[Identify the project director and collaborators who would work on the project during the proposed grant period, and describe their responsibilities and qualifications. Provide résumés for the principal collaborators (maximum of two pages each) in an appendix. Project directors must devote a significant portion of their time to their projects. All persons directly involved in the conduct of the proposed project--whether or not their salaries are paid from grant funds--should be listed, their anticipated commitments of time should be indicated, and the reasons for and nature of their collaboration explained. If the project has an advisory board, provide a statement of its function and a list of board members.]

Project Leaders (see resumés in §6.2)	
Bishop, Thomas E.	Wenlin Institute, Inc. (CEO and lead programmer); ABC Dictionary Project, University of Hawaii
Mr. Bishop is the inventor of CDL, and the principal software architect of Wenlin Institute's C programming language implementation of CDL. A staunch proponent of Unicode and educational software development, his role is key in the proposed project. Mr. Bishop will implement further refinements in the C source code, in order to produce the client application. He will also work to document the system, and introduce it to end-users. Total time commitment will be 33%.	
Cook, Richard S.	Post-doctoral research fellow, Dept. of Linguistics, University of California, Berkeley; Artificial Intelligence Group, International Computer Science Institute; Unicode Consortium
Dr. Cook is the main user and principal evangelist of CDL in the international encoding community. He is one of the editors of the *Unicode Standard*, and a maintainer of the *UniHan* component of the *Unicode Character Database*. In the proposed CDL project, he will develop the server-side software components. He will also assist in preparing the system documentation, and introduce it to end-users. Total time commitment will be 33%.	
Advisory Board Members	
Anderson, Deborah	Researcher, University of California, Berkeley Dept. of Linguistics; Script Encoding Initiative
Dr. Anderson has contributed logistical support in organizing IRG meetings, and it is anticipated that she will continue in this capacity in the future.	
Jenkins, John	Technical Director, Unicode Consortium, Apple Computer, Inc.
Mr. Jenkins is co-editor of *UniHan* with Dr. Cook, and a user of CDL.	
Lu Chin	Hong Kong Polytechnic University; Rapporteur, ISO/IEC JTC1/SC2/WG2/IRG
Dr. Lu is the leader of the IRG, and has been key in introducing CDL to IRG member delegates.	
Lunde, Ken	Adobe Systems, Inc.
Dr. Lunde is a strong proponent of CDL, and has organized many meetings on CDL at Adobe , Inc.; his contribution is essential in promoting the use of CDL in the international community.	
McGowan, Richard	Vice President, Unicode Consortium; Script Encoding Initiative
Mr. McGowan provides technical and other assistance, via the Unicode Consortium.	
Whistler, Kenneth	Technical Director, Unicode Consortium, Sybase, Inc.
Dr. Whistler is the inventor of the name "CDL", though he disagrees on what CDL stands for. Whistler will continue to assist us in the international encoding arena.	

4.4) CDL Methods

[Explain the project's methods. • Describe in detail the tasks to be undertaken and the computer technology to be employed, indicating what technical and staff resources will be required, as well as the staff's experience with the technology and its application to the humanities. • Describe plans for evaluating the results of the start-up activities. This evaluation should be simultaneously summative with regard to the Digital Humanities Start-Up Grant and formative with regard to the long-term project goals.]

The CDL XML application as detailed in the *CDL Specification* (Bishop & Cook 2003; see the links given above in §4.2) is already fully implemented in C programming language in Wenlin Institute's *Wenlin* software package, and has been thoroughly tested in the creation of more than 56,000 descriptions. This software combines a full-featured Unicode text editor with an editing interface for the Wenlin CDL font database. Wenlin's CDL font technology is used internally in the editor as one means of character display (conventional fonts may also be used). CDL descriptions define script elements which may be used recursively in the definition of other script elements. The CDL font database is simultaneously a stroked font able to display characters at any resolution, and also an index into the CDL descriptions. Thus, the same information that displays the characters on the computer screen provides access to the characters for input (hand-writing recognition or stroke-based input method) or indexing (by stroke type and component structure). CDL descriptions can serve as the basis for regular expression searches, and keys of various types can be generated from this data, in order to perform validations prior to augmentation of the set (to prevent duplication, and to regularize descriptions). The descriptions in the CDL font database are rendered by means of the Wenlin interpreter, which is built into the editor. The CDL font format also supports embedded SVG, for display in applications lacking Wenlin's interpreter.

For the purpose of the present project, the Wenlin C implementation has been augmented with open-source code (libcurl) for communication with the server-side MySQL database via an intermediary CGI application running on the server. The Wenlin application running on the client will transmit CDL descriptions to the server, where they will be stored with versioning and other information, in the MySQL database. The Wenlin client will communicate user access control information to the server, and all changes to the centralized server database will be associated with a specific user. The CDL client at present allows practically infinite roll-back, via a variant mechanism which supports up to four billion CDL descriptions per Unicode code point. Similar roll-back functionality will also be supported in the server database. Very important in the long-term, CDL's variant mechanism is key for addressing the important issue of distinctive CJK character variation (in conjunction with variant selectors, registries, etc.).

The prototype client to be prepared in the current proposal will be a proof of concept, and we will not try to implement at once all of the features that will eventually be a part of the system. Instead, the client applications will have rather limited capabilities, aimed primarily at completing descriptions of the remaining 16,000 encoded characters. A future step will involve creating descriptions of the tens of thousands of characters which are currently in the pipeline for future Unicode encoding. This latter process can be boot-strapped on the basis of componential descriptions which are currently being constructed by other IRG contributors. A similar boot-strapping process was already employed in the creation of the existing CDL descriptions.

The CDL font editor is rather easy to use, but at present the only documentation is that available on the CDL website. In order to introduce collaborators to the use of CDL, tutorials will be conducted at IRG meetings during the funding period. The CDL team has already presented CDL at several IRG meetings, and both project leaders have extensive teaching experience, which should provide for a gentle introduction to this powerful technology.

4.5) Final CDL Product and Dissemination

[Describe the plans to disseminate the project results through various media (printed articles or books, presentations at meetings, electronic media or some combination). Applicants should also discuss how the project's ultimate product is likely to be disseminated and what provisions will be made for the long-term maintenance of such a product.]

The deliverables resulting from this CDL project will be functional software prototypes, documentation, and increased awareness of CDL in the international encoding community. Software prototypes of both the client and server will be produced, and made available to IRG delegates. The CDL documentation on the current CDL website will be expanded (leading eventually to a book on CDL, though the book might not be completed in the funding period). The elements of the CDL system and documentation will be presented at IRG conferences, to introduce details of system usage. The public CDL data repository will also have a simple browser-based query interface, for regular expression searches, and this will also be available on the CDL website at the conclusion of the funding period.

4.6) CDL Work Plan

[Describe the specific tasks that will be accomplished during the grant period and identify the staff members involved. The start-up activities described in the proposal should be completed by the end of the grant period.]

The Work Plan for the CDL start-up has sub-projects relating to accomplishing three main tasks:

1.) **Modifications to the Wenlin source code:**
Mr. Bishop will undertake modifications to Wenlin Institute's existing code-base, to permit the Wenlin client application (which runs on Windows and Macintosh computers) to communicate with the server-side MySQL database.

2.) **Server-side database and CGI application programming:**
Dr. Cook will create the server-side database environment (MySQL database and CGI programs) for communication with the Wenlin clients, and for the browser-based query system. These components will be hosted on one of the many servers on the domains at the CDL team's disposal (*wenlin.com*, *linguistics.berkeley.edu*, or *unicode.org*).

3.) **Documentation and promotion:**
The current CDL website (<http://www.wenlin.com/cdl/>) will be augmented with additional documentation, such as is necessary for end-users to understand proper usage of the CDL client. This website will also be updated with links to the browser-based query system, and with status pages giving status on the CDL MySQL database.

Work on these three components of the project will be undertaken concurrently, over the 18-month funding period, with a 33% time commitment by each of the CDL project leaders. We anticipate that prototypes of the two main software components, and also of the browser-based interface for regular-expression searching, can be completed within the first twelve months of funding, and that during the end of that period and in the remaining six-month funding period the system can be refined and made available to IRG members for initial testing.

The many special functions of the Wenlin client application (Unicode text editor, font editor, dictionary interface) will ideally be available in the future in a browser-based version, circumventing the need to install the stand-alone Wenlin client on individual machines. Creation of a browser-based version of Wenlin is however a more distant goal, dependent upon exploration of the necessary technologies.

5.0) CDL Project budget

[Budget narrative (optional) If needed, include a brief supplement to the narrative explaining projected expenses or other items in the financial information provided on NEH's budget form. The budget narrative may be single-spaced.]

The CDL project budget sheets are attached on the following pages, seeking funding for the 18-month period beginning in April, 2007.

6.0) CDL Project Application Appendices

[Use appendices to provide essential supplementary materials. Include a brief résumé (two-page maximum) for each principal project participant and letters of commitment from other participants and cooperating institutions. Descriptive material from preliminary work or previous periods of support may be included in an appendix, but should be limited to essential information.]

The elements of the CDL system were first publicly specified in two core documents (Bishop & Cook, 2003; these are included in this appendix, for the convenience of reviewers reading paper print-outs of this application):

- **"A Specification for CDL: Character Description Language"** .
  <http://www.wenlin.com/cdl/cdl_spec_2003_10_31.pdf >
- **"Character Description Language (CDL): The Set of Basic CJK Unified Stroke Types"**
  <http://www.wenlin.com/cdl/cdl_strokes_2004_05_23.pdf>

Additionally, we attach a copy of the recently approved encoding proposal **N3063**, adding 20 stroke types to the **CJK Strokes** block of Unicode (ISO/IEC 10646):

- **"Proposed additions to the CJK Strokes block of the UCS"** .
  <http://www.dkuug.dk/jtc1/sc2/wg2/docs/n3063.pdf>

The CJK Strokes block, encoding a total 36 strokes (U+31C0 .. U+31E3) resulted from the proposers' work with the US and international standards bodies (Unicode and ISO/IEC 10646/WG2/IRG), and the proposal itself was completed in IRG sub-committee work hosted at UC Berkeley in November, 2005 (with financial assistance from the UC Berkeley Townsend Center for the Humanities, the UC Berkeley Dept. of Linguistics, the UC Berkeley Institute of East Asian Studies, and the UC Berkeley East Asian Library).

For more information on the domestic and international standards bodies and other organizations contributing to this work, please see the following links:

- **The Unicode Consortium**
  <http://www.unicode.org/>
- **The Ideographic Rapporteur Group (ISO/IEC 10646/WG2/IRG)**
  <http://www.cse.cuhk.edu.hk/~irg/>
  <http://www.cse.cuhk.edu.hk/~irg/irg/irg25/IRG25.htm>
- **The Script Encoding Initiative**
  <http://www.linguistics.berkeley.edu/sei/>
- **The Sino-Tibetan Etymological Dictionary and Thesaurus Project**
  <http://stedt.berkeley.edu/>
- **Wenlin Institute, Inc.**
  <http://www.wenlin.com/>

The final two appendices include resumés for the CDL project leaders.