

Introduction

Digital libraries offer the potential of anytime, anywhere access to a diversity of knowledge. Understandably, this promise has attracted attention and resources. While these digital library efforts are noteworthy, the reality that vast amounts of knowledge contained in print format remains untapped in these digital environments is disconcerting. The PITAC report on digital libraries asserts that “barely 10 percent of public information in print has been digitized and made available on the Internet” (PITAC 2001). With increasing reliance on digital resources, efficient ingestion of print materials into digital libraries becomes increasingly important. The ultimate realization of digital libraries should include knowledge from all aspects of human scholarship and unleash the full potential of multiple media including image, text and sound. Greater ingestion of print materials, especially with processes designed to create multiple media, will help populate digital libraries with a richer array of content and services that is more representative of human knowledge. But constraints in available technology and resources often inhibit data capture of materials related to the humanities.

In this regard, the materials and research methods of humanists present fundamental challenges for information technology research. The hardware and software used for data capture of “standard” documents (e.g., business documents, modern paper documents) is often inadequate for ingestion of humanities materials. To this end, Johns Hopkins University proposes to lead an effort to build a framework, consisting of inter-linked hardware and software components, which will enable the efficient ingestion of humanities-based materials. The Digital Knowledge Center, led by PI Choudhury at Johns Hopkins, has focused on the creation of automated tools, processes, and systems to facilitate the ingestion of print materials while reducing the amount of human labor required for these efforts. This research is founded on the principle that while human labor will be required for data capture, manual intervention should be required only for more focused and highly skilled tasks, rather than for mundane and repetitive ones. That is, the combination of automated technology and, when necessary, human domain expertise results in more efficient, practical, cost-effective and scalable systems for data capture and ingestion for digital libraries.

The proposed research will result in a fully automated robotic system for on-demand and batch scanning of print materials (“CAPM”) and an open-source software framework for document analysis that can be trained and calibrated by humanists (“GAMERA”). The resulting system will include an inter-linked mechanism between CAPM and GAMERA. To evaluate different techniques for document analysis, including GAMERA, we will build a testbed of digital images. GAMERA will be usable—designed according to the principles of usability which include effectiveness, efficiency and satisfaction.

The project team features an international multidisciplinary team of humanities faculty and post-docs, librarians, a usability specialist, and interdisciplinary digital library researchers. The personnel represent Johns Hopkins University, Tufts University, Edinburgh University, and the University of Oxford. Additionally, the British Library has agreed to provide resources for this project. This combination of diverse disciplinary expertise reflects the complex and demanding facets of conducting the proposed research. These institutional partnerships, which have evolved naturally due to common interests and goals, represent the culmination of existing collaboration. The investigators and senior personnel will build upon successful results from existing research and will advance these efforts in substantive and fundamental ways through ITR funding.

Why the Humanities?

Although the program solicitation for the ITR program cites specifically the potential role of the humanities, it is worthwhile to consider why the perspective and needs of humanists, especially as they relate to data capture, are relevant in the context of information technology research. Aside from the fact that humanists are an underrepresented group within the realm of information technology research, there are three major reasons why more efficient data capture of humanities-based materials is important. First, it is important to include humanities content to ensure full representation of human knowledge in digital libraries. Second, greater availability of humanities digital content could enable new modes of inquiry for humanists. Finally, and perhaps

most importantly, extracting knowledge from humanities content poses challenges that will foster innovative information technology research. Each of these reasons is discussed below.

(1) With any library, traditional or digital, it is essential that a broad spectrum of human knowledge be represented. Excepting special libraries, one would most probably consider a library that contains only scientific documents or only medieval French documents to be incomplete. Within the context of digital libraries, successful efforts such as the Perseus Project, led by co-PI Crane, have provided a foundation for creating digital content related to the humanities (Crane 1998). However, the experiences of co-PI Crane's group and PI Choudhury's group have also led to the realization that data capture of humanities materials presents significant challenges and raises questions regarding the scalability of current technologies and methodologies. Without a significant improvement of data capture procedures, it is possible that digital libraries will develop with a disproportionate emphasis on materials that are more easily ingested or "born digitally" rather than an emphasis on intellectual needs. Libraries do not acquire the cheapest or most easily obtainable materials; rather, they acquire materials driven by the intellectual needs of their scholars and patrons.

(2) Greater ingestion of humanities materials could change the methods of humanists' inquiry and research in fundamental ways. Specifically, a transition from "data poor" to "data rich" environments could lead to new modes of inquiry, research, and instruction for the humanities. The Perseus Project has provided evidence in this regard (Crane et al. 2001a; Crane et al. 2001b). The ingestion of digital images from humanities materials and extraction of text and musical content will increase the volume of content available for research and inquiry and will make this content more readily available. Basic applications of information technology within the humanities have yielded benefits. Information technology allows humanists to write and revise articles more efficiently in document management systems instead of on typewriters; they download and print articles not available in local libraries. Additional information technology research could lead to an increase of content and associated services for research and instruction that will have profound impact on humanities scholarship. In this regard, Co-PI Nichols has focused on the specific case of medieval studies.

Over the last two decades, medieval studies have turned more and more towards studying manuscripts as artifacts important not simply for their content, but as unique sources of information about medieval society in their own right. Recent studies of manuscripts have looked, for instance, at the nature of the materials contained in them – why *this* particular configuration of texts? They have also noted historical references introduced in one particular version of a well-known text not contained in other manuscripts. Sociological implications of the "book" trade have been tracked, such as the different qualities of manuscripts produced and for whom, information that allows scholars to trace the relative affluence of different levels of society in given cities at a particular period. Finally, by correlating interpolated comments in manuscripts with the place and time of their production it has been possible to surmise some of the issues people thought about at a given time and place which were different from issues of the moment elsewhere at the same time.

Many more historical and sociological insights have been gleaned from the new approach to manuscript study. The point, however, is not the riches awaiting discovery in these artifacts, but the ways in which we might make these essential historical documents accessible. When manuscripts were primarily of concern to editors preparing editions of a given text, or to historians looking for archival information, the problem of access was less crucial because it was not a quantitative one. A few scholars dealt with a given set of manuscripts of canonical works. They typically spent years, even a whole career, poring over the small number of manuscripts that would constitute their edition. Critical editors were far less interested in a given manuscript itself than in the edited text they would ultimately produce. In a sense, the manuscript was "rare material" for their ultimate product.

All this has changed. Now medievalists are interested in studying manuscripts themselves and the information they contain. Quantitatively at least, the biggest single "chunk" of information in a manuscript is the text it contains. Since we are no longer solely interested in one, "best" version of a text, we need transcriptions of the text from a number of different manuscripts for comparative purposes. That is the single most time-consuming process. It is also, in way, a rote task, even though one that requires great technical skill.

One can be an expert on one manuscript, or a set of manuscripts all produced in the same scriptorium, but that does not mean that one can pick up another manuscript and begin to read it straight away. A different hand will require a learning process, perhaps half a day for the clearest bits, but longer for the passages where the letters are formed differently, or where something may have been omitted, or where a word has been corrupted. Those are the passages that test one's expertise. They are the passages that we may never be able to resolve by optical symbol recognition. If one could create optical symbol recognition software that would read the majority of the text, however, it would make a greater number of texts available at lower cost (paying for transcribers to copy manuscript by hand is enormously expensive). It would also free up experts from the rote tasks of transcribing so that they could then focus on resolving the difficult bits which exist in every manuscript and which sometimes require hours or days to decipher. A small group of scholars would then be able to undertake in depth study of a large number of manuscripts, the advantage being that the study group would be able to formulate a coherent set of questions that could be applied to the whole corpus of manuscripts. Medievalists have never before been able to do this for a corpus consisting of more than ten or a dozen manuscripts. With the proposed work, it would now be possible to do this for works having many dozens of manuscripts.

(3) The final, and perhaps most important, reason for considering the humanities in the context of information technology research focuses on the challenges presented by extracting knowledge from humanities materials. The physical nature and diverse content of humanities texts, manuscripts, maps, and other documents pose challenges for information technology research. For example, even the best commercially available optical character recognition (OCR) software cannot be trained to handle pre-nineteenth century typefaces. Additionally, character recognition is only a basic step in the overall process of data conversion. To be useful, for example, a number like *1922* must not only be recognized as a sequence of digits on a page. It must be recognized as occurring in a particular context on the page—in the leftmost element of a table row, perhaps—and ideally its semantic function should be recognized as well. This same concept of analyzing layers of information applies to musical symbols that must be recognized and interpreted as more than pixels on an image.

To deal with these issues, some researchers have developed techniques for specific elements or specific types of humanities content such as cursive handwriting in manuscripts (Cerquiglini et al. 1998; Keaton et al. 1997; Manmatha and Croft 1997). Other researchers have developed frameworks to accommodate a class of documents such as technical documents (Thibadeau et al. 1995). The most robust system, however, should account for the broadest range of image processing problems. Arguably, the universe of humanities documents provides the most difficult challenges in this regard. The proposed research will result in a framework that accounts for and a testbed that reflects this diversity of challenges.

Beyond the implications for information technology research of data capture from the diversity of humanities content, humanists also pose inquiries that identify limitations of technology such as OCR. Humanists are often interested in specific symbols or specific patterns within documents. During a recent visit to the University of Edinburgh, PI Choudhury discussed a representative case of this mode of inquiry with project personnel Laidlaw and Ovenden. Professor Laidlaw is interested in determining which portions of Christine de Pisan manuscripts were written in her hand. For this specific purpose, the ability to measure similarity between handwriting of passages is more relevant than identifying specific words. Additionally, Johns Hopkins has initiated a project to digitize, classify, and encode cuneiform from clay tablets. With cuneiform (and, for example, hieroglyphics), it is necessary to recognize unique symbols that pose problems for OCR technology.

Printed texts, though much more tractable than manuscripts, still pose major challenges for OCR technology. The best commercial OCR systems combine and average the results of multiple OCR engines, but this voting system means that the systems are optimized for the least common denominator content areas, primarily business documents in English, with reasonable support for other Western European languages in Roman alphabets. Furthermore, high-end systems with multiple embedded OCR engines are often difficult to train and extend. While individual trainable OCR engines often assume a core set of characters. Thus an English OCR engine will not be able to learn Greek, and even a Cyrillic OCR engine may not be sufficiently flexible to learn

Greek accents. OCR engines with fixed (and invisible) prototypes may not be able to learn a trivial typographic feature such as the “long s” (an “s” that closely resembles an “f”) in earlier printed books. Even relatively modern print cultural heritage materials can be problematic. Newspapers constitute one of our major records for social and political history, but many newspapers now exist only in microfilm surrogates of uneven quality. Humanists need to be able to tune OCR systems to optimize their ability to produce searchable text from such degraded sources. For these types of inquiries, a framework that enables robust document analysis is necessary. The proposed research will further develop an existing image-processing framework, GAMERA, which will accommodate the diversity of challenges presented by humanists’ documents and research questions.

To continue the development of GAMERA, and to explore the aforementioned issues, it is necessary to have a testbed of digital images from humanities collections. We will evaluate GAMERA (and other available document analysis methodologies) using this testbed, which will be developed with the specific goal of expanding the diversity of image processing problems presented by the various forms of humanities content. The project team has taken initial steps toward the development of this testbed. Specifically, existing humanities digital collections with associated transcriptions provide the foundation for this testbed. Portions of additional rare humanities collections will be digitized to augment this testbed. These collections will be chosen specifically to enhance the diversity of the testbed (and thereby provide a greater range of image processing problems). This testbed development will be consistent with the digital library development at the participating institutions and will be possible given formal partnership agreement between Johns Hopkins and the University of Oxford and a similar agreement between Johns Hopkins and Edinburgh University. Normally, Oxford and Edinburgh would not allow access to their digital collections in the proposed manner.

Additionally, an existing robotic system for on-demand and batch scanning of print materials, CAPM, will be further developed to generate continuous, automatic additions to the testbed. With a large and diverse testbed of digital images of cultural heritage materials, it becomes possible to extend the GAMERA framework most effectively to address the greatest range of image processing problems.

The next section describes the existing foundation, including current digital collections, and the current status of both CAPM and GAMERA. The subsequent section describes the additional digitization of humanities collections, and the extension of CAPM and GAMERA that ITR funding would enable.

Existing Foundation

The foundation for the proposed work comprises existing digital collections that have been used to develop the core of GAMERA and the CAPM robotic system for automatic scanning of printed materials in remote locations. Typically, rights management guidelines at the participating institutions constrain the use of these digital collections for research and development efforts. However, for this proposal, the participating institutions have agreed to provide digital collections for refinement and calibration of GAMERA. The next subsections describe the existing testbed and the current status of CAPM and GAMERA.

Digital Collections

Lester S. Levy Collection of Sheet Music

The Lester S. Levy Collection of Sheet Music comprises nearly 30,000 pieces, which corresponds to nearly 130,000 sheets of music, with an emphasis on popular American music from 1820 to 1914. The Collection provides a snapshot of American society through its cover art, music and lyrics. The digitized collection serves as the foundation for Johns Hopkins’ NSF DLI-2 project (“Levy II”) to develop a digital workflow management system (Choudhury et al. 2000). This workflow management system features an automated metadata generation tool (DiLauro et al. 2001), a disk-based search engine (Droettboom et al. 2001) and optical music recognition (OMR) software (Choudhury et al. 2001a). OMR provides the foundation for the more generalized GAMERA. The music information retrieval (MIR) community has initiated an effort to build a testbed for MIR research. The digital Levy Collection has been identified as a potential component of this effort.

Electronic Corpus of Lute Music

Professor Tim Crawford from King's College has given us an electronic corpus of lute music.

This collection of lute tablature contains: J-B. Besard's *Thesaurus Harmonicus* printed in 1603 (185 pages); Cambridge University Library MS Dd.2.11 (one of the most important manuscripts of the Dowland/Shakespearean period, c. 1600) (100 pages); British Library MS Add. 31392, which begins with MS keyboard music in normal notation and in tablature (48 pages); and Simone Molinaro's *Intavolatura di Liuto* of 1599, which is in Italian tablature (75 pages). This collection will enable further refinement and development of OMR.

Roman de la Rose

Co-PI Nichols and Senior Personnel Requardt have led a project to test ways to present medieval manuscripts in digital form. The project features digitized images from three manuscripts of the *Roman de la Rose* text from the collections of the Walters Art Museum (W. 143), the Pierpont Morgan Library (M. 948) and the Bodleian Library of Oxford University (MS. Douce 195). Medieval French scholars, led by co-PI Nichols, have transcribed the existing three scanned manuscripts. This combination of images and transcribed texts represents an excellent opportunity to test and calibrate GAMERA for recognizing medieval French.

Perseus Digital Library

The Perseus Digital Library contains collections from antiquity through the twentieth century. The Perseus team has carefully edited transcriptions of many different types — including non-English, non-Roman documents such as Greek source texts, non-English, pre-modern such as Latin, and non-standard modern language documents (such as earlier English books with long s's, German Fraktur). The carefully edited transcriptions can provide both training sets and tools for automatic evaluation of optical symbol recognition results.

Statistical Accounts of Scotland

The provision of an online version of the *Statistical Accounts of Scotland* has provided a major national resource available free of charge worldwide. The initial project however, relied on an OCR version of the digital page-images which was completed several years ago. The original text was printed in 18th and early nineteenth century typefaces, using letter-forms and ligatures which were typical of the day, but which posed serious problems for the original OCR project. Being able to subject the page-images to more sophisticated optical recognition software raises the potential of supplying a much cleaner underlying text to enhance search and retrieval techniques on this significant online resource.

Oxford University Library Services

The University of Oxford's many libraries contain the largest and most diverse collections for the support of teaching and research in any institution of higher education in the United Kingdom. Its library holdings are world-class. Its principal library, the Bodleian, has been in effect a library of legal deposit for almost 400 years. The libraries that comprise Oxford University Library Services (OULS) contain about 10 million volumes. A large proportion of the library stock will be preserved in perpetuity and a significant percentage of copyright revenue will form part of the national printed archive.

In order to improve access, OULS has recently embarked on an ambitious program to create the Oxford Digital Library. GAMERA will be presented with a range of document analysis problems presented by the following collections that have been digitized as part of the Oxford Digital Library.

Celtic and Medieval Manuscripts Project (<http://www.image.ox.ac.uk>)

An extensive collection of very high quality images of medieval manuscripts held at the Bodleian and six Oxford college libraries has been digitized since 1995. This project began by digitizing manuscripts of Celtic origin, but has since been expanded to incorporate a range of medieval material, and also a small number of Ancient Roman papyri.

Internet Library of Early Journals (ILEJ) (<http://www.bodley.ox.ac.uk/ilej/>)

A collaborative project between the universities of Oxford, Leeds, Birmingham, and Manchester. It chose six journals – *The Builder*, *Notes & Queries*, *Blackwood's* (nineteenth century), *Philosophical Transactions of the Royal Society*, *Gentleman's Magazine*, and the *Annual Register* (eighteenth century) – covering a ten or twenty year run from each, with the total images amounting to around 108,000.

Bodleian Library Broadside Ballads (<http://www.bodley.ox.ac.uk/ballads>)

The Bodleian Library has unparalleled holdings of over 30,000 ballads in several major collections. The original printed materials range from the 16th to the 20th Century. The Broadside Ballads project makes the digitized copies of the sheets and ballads available to the research community.

John Johnson Collection of Printed Ephemera (<http://www.bodley.ox.ac.uk/toyota>, http://www.ilrt.bris.ac.uk/jidi/col_john.html and <http://www.bodley.ox.ac.uk/johnson/>)

The Toyota City/Bodleian Library Imaging Project was Oxford libraries' first completed digital imaging project. The project produced 7000 scanned images of transport ephemera from the John Johnson collection. Currently, political prints, cartoons and advertising (approximately 2500 items) are being scanned as part of the JISC Image Digitization Initiative (a major national endeavor which has produced 50,000 high quality digital images over 2 years), and its collection of trade cards is being digitized as part of a separate project.

Browsable Catalogue of Medieval Manuscripts

(<http://www.bodley.ox.ac.uk/dept/scwmss/wmss/medieval/browse.htm>)

Approximately 700 images of medieval and renaissance manuscripts, most of which have been catalogued as part of a project to produce the first catalogue of about 500 manuscripts acquired by the Bodleian since 1916, have been digitized during 2000. The manuscripts range from the 12th-17th centuries. Many of the images are being drawn from the Bodleian collection of filmstrips, which contains c. 30,000 frequently requested images.

Early English Books Online - Text Creation Partnership (<http://www.lib.umich.edu/eebo/>, <http://www.uni.com/eebo/>)

Oxford University is a partner in the EEBO (Early English Books Online) project, which has scanned all 96,000 titles catalogued in Pollard and Redgrave's Short Title Catalogue (1475-1640), Wing's Short Title Catalogue (1641-1700) and the Thomason Tracts (1641-1660). The University is participating in the next stage of the project, a Text Creation Partnership with University of Michigan which aims to make a proportion (c. 20-25%) of these texts available in fully searchable and encoded form.

Comprehensive Access to Printed Materials (CAPM)

The Comprehensive Access to Printed Materials (CAPM) project began in response to a pressing problem facing many libraries, especially academic research libraries. Given the increase of electronic resources (and associated infrastructure) within libraries and ongoing print-based acquisitions, most libraries face major space shortages (Wagner 1995). In response to this space constraint, libraries have often built large, off-site shelving facilities to house portions of their collections. Johns Hopkins University has implemented such a facility named Moravia Park. The Moravia Park facility offers high-density shelving with individual books arranged by similar size and contained within open boxes. While this strategy addresses the space problem, it eliminates the ability to browse for materials shelved in these off-site facilities.

The CAPM Project was initiated to address this lack of browsability. With initial funding from the Council on Library and Information Resources (CLIR), PI Choudhury conducted a technical and economic feasibility analysis. This initial analysis provided evidence that a robotic system could be developed to restore browsability to printed materials shelved in off-site facilities. The system itself will operate as follows: when a patron requests an item that is shelved off-site, a retrieval robot will automatically navigate the stacks to find the desired item and bring it to a scanning station. The item is then transferred to another robotic system that is responsible for scanning. OCR software (or GAMERA) would subsequently process the images. Upon completion of the scanning the retrieval robot will return the book to the stacks or separate the book for physical

delivery to the main library building. It is important to note that CAPM differs from existing warehouse automated retrieval systems within libraries (Hansson 1995; <http://library.csun.edu>) since it is a on-demand and batch *scanning* system for printed materials in remote locations. Ultimately, the CAPM system will result in the ability for patrons, even outside of Johns Hopkins University, to browse materials shelved at the off-site facility independent of space and time making cost-effective on-demand digitization a reality.

Following this initial feasibility study, the Mellon Foundation funded the first phase of the CAPM Project. This first phase of CAPM resulted in the development of a prototype retrieval robot and a concurrent economic cost-benefit analysis (Choudhury et al. 2001b; Suthakorn et al. 2002). The CAPM system differs from existing systems in the following ways. First, the system retrieves individual items, as opposed to boxes of items. Second, the CAPM system does not assume an existing or fixed shelving and space arrangement. This flexibility will allow it to work in many diverse environments. Third, the CAPM retrieval robot is autonomous and can navigate independently (i.e., it does not require remote control). Fourth, the economic analysis has provided evidence that the relatively inexpensive system is cost-effective, especially in comparison to potential benefits (<http://dkc.mse.jhu.edu/CAPM>). Finally, the page-turning system, to be built with NSF ITR funding, will accommodate a wide diversity of paper types and materials.

Even though CAPM was initiated with a goal of restoring browsability for printed materials at remote locations, the system has another benefit that is more relevant for the present discussion. Once the robotic page-turner is operational and CAPM is not busy with a patron request, the system can scan materials continuously and automatically. This ongoing digitization of humanities materials will contribute to the testbed used to test GAMERA. The overall testbed development effort (and associated budget request) reflects the assumption that CAPM will become the primary mode for digitization of cultural heritage materials.

Since CAPM will be used for a collection of materials selected by the Libraries at Johns Hopkins (and other institutions who have housed materials at Moravia Park), the scanning of materials will reflect the intellectual organization and coherence of the libraries' collections. For this reason, some portion of the collection will contain cultural heritage materials that are consistent with the "collection development" strategy of the GAMERA testbed.

GAMERA

From the beginning of the aforementioned Levy II Project, one of the most important goals was the development of a flexible optical music recognition system, capable of converting the historical sheet music in the Levy Collection into a symbolic format. To achieve this goal, an adaptive OMR system developed by co-PI Fujinaga of the Peabody Conservatory of Music at the Johns Hopkins University (Fujinaga 1997), was chosen as the basis for the Levy Project software and expanded with an Optical Music Interpretation system (Droettboom and Fujinaga 2001). OMR is not sufficient to fully analyze the documents in the Levy Collection, however; text is present as score markings, lyrics, and metadata. It was hoped that the text recognition needs of the project could be met with an existing optical character recognition (OCR) system. Unfortunately, testing revealed that this was not a viable solution. The OCR systems tested either performed poorly or were not suitable for the batch-processing environment necessitated by the size of the Levy Collection.

The failure to find an existing OCR system appropriate for the Levy project, alerted us to the general need for flexible document analysis tools suitable for the creation of content for digital libraries. In particular, document analysis tools are generally not available for documents in ancient languages that contain non-standard printing, or for a variety of reasons differ from common documents. This lack of tools specifically targeted at these types of documents is not likely to change in the future because of the limited market for them. Many of these problems are similar to those solved by the OMR system, however. At the core of the Levy OMR system is a general symbol recognition system capable of learning new symbols. In order to address the OCR needs of the Levy II Project and other future projects it was decided to generalize the existing OMR technology to make it suitable for a variety of document analysis tasks, including text and music recognition.

The resulting system, called GAMERA, is a toolkit for the creation of domain-specific document analysis systems by document experts. The overall design is inspired by systems like MathWorks Matlab and CVIP tools (Umbaugh 1998). GAMERA is implemented as a set of extensions to Python written in C++ and Python. The goal is to leverage the user's knowledge of the target documents to create custom applications rather than attempting to meet the needs of diverse users with a monolithic application. GAMERA is a graphical environment that allows users without extensive programming experience to be productive with a minimum amount of training. The components provided in GAMERA can be combined to create complete document analysis applications, but additional tools and approaches will be useful for certain document types. To address this need, GAMERA includes a system for the easy creation of plug-ins by third parties in either C++ or Python.

The specific features of GAMERA are:

1. Graphical environment for the creation of custom document analysis applications.
2. Flexible tools including a learning symbol recognition system.
3. Rich user interface components for development and training.
4. Suitable for use in large-scale digitization projects.
5. Extensible by third parties with Python and C++.
6. Open-source and standards-based for easy integration with workflow management systems.

The user interface elements provided by GAMERA are used to create a development environment and provide an efficient interface for training the learning classifier. The creation of an effective user interface was one of the most time-consuming aspects of developing GAMERA, but it has proven to be the most useful. The wxPython library (<http://www.wxpython.org>) was chosen as the toolkit for the user interface because of its portability (Windows and Unix) and the completeness and flexibility of its widgets (MacMillan et al. 2001, MacMillan et al. 2002).

Although the design of GAMERA is similar to systems like MathWorks Matlab, CVIP tools (Umbaugh 1998), Feature Center (Thibadeau et al. 1995), and HUE (Cracknell and Downton 1999), it differs from these systems. Many of these systems are not specifically designed for document analysis. While Feature Center is intended for document analysis, its domain of technical documents is less diverse than the intended domain (of cultural heritage materials) for GAMERA. Perhaps, the "closest" comparison can be made between HUE and GAMERA. However, there are some important differences.

HUE and GAMERA are both structured-document analysis application development environments with similar goals and philosophies. GAMERA, which is targeted at a less technically sophisticated audience, includes many important features that allow it to be more extensible and easy-to-use. Specifically, GAMERA has a modern object-oriented design and a more flexible component model. Additionally, GAMERA has an extensive GUI that includes a flexible training and ground-truth creation interface suitable for use by non-programmers. Many of the advantages of GAMERA are realized through the use of the more modern object-oriented languages C++ and Python. By contrast, C and Tcl/Tk constrain HUE.

Proposed Work

This section outlines the activities that would be supported by NSF funding. The coordinated activities are intended to maximize the development of document analysis capabilities offered through GAMERA. Once CAPM is deployed, it will become the primary method for digitization of cultural heritage materials for the testbed. However, there are certain materials, such as rare manuscripts, that would not be housed within an off-site facility like Moravia Park. These materials will increase the diversity of humanities content. Consequently, digitization of these materials will take place independently of CAPM during the first years of this project. The budget request reflects the shift from manual digitization of rare materials to automatic, continuous batch scanning through CAPM. The following subsection describes the digital collections to be digitized and included within the GAMERA testbed.

Additional Special Collections

Fowler Collection

The Fowler Collection, one of the premier collections at Johns Hopkins, contains numerous editions of the classic architectural treatises. It is particularly rich in the works of the Roman architect Pollio Marcus Vitruvius and the five great protagonists of the Italian Renaissance Andrea Palladio, Leon Battista Alberti, Vincenzo Scamozzi, Sebastiano Serlio, and Giacomo Barozzi Vignola. We intend to digitize fifteen of their works. GAMERA would benefit from the use of these early architectural treatises as test cases for two reasons. First we would be able to extend the OCR capabilities to the typefaces used in early printed books. Second, we hope to be able work on the symbolic recognition of architectural diagrams, which have heretofore either been treated as simple images or laboriously hand-referenced by a trained person with a digitizing tablet.

Emblem Books

Johns Hopkins is also planning on digitizing a number of emblem books from the late sixteenth and early 17th centuries. Emblem books are illustrated works used for meditation. The emblem was a woodcut or engraving with the meaning of the moral lesson in the image interpreted by an accompanying motto, epigram, verse or prose explanation. The 10 works we have selected are particularly complex examples containing a variety of fonts and languages. For example in the *Imago Primi Saeculi Societatis Jesu* (1640) each chapter text is in Latin, interpreted by a poem set in italic type, and followed by the emblem image. Among the text passages can be found Greek and Hebrew phrases which include the vocalization marks. Each page has glosses (or referential text) set in the margins. These works provide a particular interesting problem for GAMERA not only because of their extraordinarily rich typography, but also because of the structural complexity of the work, which can include interrelated text in multiple languages, glosses, emblems and other figures, and extended commentaries. Since the symbolic information provided by GAMERA might include positional information, it is possible that this information could be provided to a system for automated structural metadata generation.

MS 19 Bible Historiale

MS 19 in Edinburgh University Library is an important example of the genre of historical religious works, based on the *Historia Scholastica* of Peter Comestor, in this instance in the French version of Guiart des Moulins. An Edinburgh merchant gave the manuscript to Edinburgh University Library in 1680. The illumination is particularly interesting, and has been the subject of some scholarly interest, notably J. Diamond, 'Manufacture and Market in Parisian Book Illumination' in *Europäische Kunst um 1300' Akten des XXV. Internationalen Kongresses für Kunstgeschichte* (1986); and J.J.G. Alexander, 'Preliminary Marginal Drawings in Medieval Manuscripts', in *Artistes, artisans et production artistique au Moyen Age* ed. X. Barral I Altet (1986-90), iii, 307-19. Contemporary scholars are divided as to the script: whether one or more scribes have been working on the manuscript, and whether the script is localisable to northern or southern France. Project GAMERA, by comparing the script with other scripts in digitized form e.g. the Christine de Pisan manuscript in the British Library will be able to contribute to this study. The British Library has agreed to provide access to the images from the Christine de Pisan manuscript for testing with GAMERA.

MS 56

MS 56 is arguably the most important book in Edinburgh University Library. A Psalter, written in the 11th century, it has long been the subject of scholarly debate. The arguments center around both the script and the decoration, which suggest to some an Irish influence, and to others the work of Scottish scribes and artists associated with the court of Queen Margaret of Scotland. If the latter proves to be the case, then the book becomes one of the oldest extant Scottish books. GAMERA would benefit from this manuscript being a test-case as the script is very different from the Latin book hands of the French manuscripts suggested elsewhere, as the Celtic Psalter is written in an insular text hand which poses numerous difficulties for any optical recognition software.

Thomas-Walker Collection

The collection comprises 2, 500 engraved portraits on paper, dating from the 15th to the 19th centuries, a

high proportion of which have been extracted from printed books. The portraits therefore contain a great deal of text in addition to the portraits themselves. GAMERA would be used to delineate the portrait from the accompanying text.

Oxford University Library Services

In addition to existing digital collections OULS, with curatorial guidance, will select appropriate material from the Bodleian Library's Oriental and Special Collections & Western Manuscripts sections for digitization in order to build additional sets of testbed material for the GAMERA project. The presence of the Oriental and Special Collections in particular introduces the potential for a class of materials not currently available in digitized form.

The Western Manuscripts section of the Oxford University Library Services (OULS) holds the second largest collection in Britain, with items ranging in date from papyri of the 3rd century B.C. to correspondence and papers of the present. Particular strengths are medieval manuscripts, 17th-century literary and historical collections, antiquarian and topographical manuscripts, and modern scholarly, literary, and political papers.

The Library has acquired Oriental printed books and manuscripts since its re-foundation in 1598 by Sir Thomas Bodley, himself a Hebraist. The Department's current acquisitions largely reflect the teaching and research undertaken in the University in Hebrew, Islamic, South Asian and Far Eastern studies, and important collections are also maintained in areas such as Central Asia, Southeast Asia and Tibet. Many of its collections have an importance that is truly international. The range of subjects covered by the Department include: Arabic, Armenian, Central Asian Studies, Chinese, Ethiopic, Georgian, Hebrew, Islamic Studies, Japanese, Korean, Persian, South Asian Studies, Southeast Asian Studies, Tibetan, Turkish.

Additionally, the Bodleian Library possesses a rich collection of maps from various periods. The application of GAMERA to digital images of maps would enable exploration of name extraction combined with recognition of roads, contours, and other continuous symbols. Another collection that has been identified is the scores, drawings and correspondence of Mendelssohn.

Proposed Development of CAPM

As mentioned previously, a fully automated CAPM system will be developed to enable continuous digitization of cultural heritage materials. The next subsections describe the issues related to development of the CAPM system.

Scientific and Technical Issues

This subsection begins with a review of the current technology of the CAPM robotic book retrieval system that was developed with support from the Mellon Foundation. We then enumerate open scientific and technological problems that will be solved during the course of the proposed work. Various feasible solutions are examined, the most promising of which will be breadboarded and compared during the first year of the proposed work. In subsequent years functioning prototypes will be built and tested based on experimental observations of breadboarded models obtained in the first year.

Review of the Current Retrieval Robot and Open Issues

The current CAPM retrieval robot consists of a six-degree-of-freedom CRS manipulator arm mounted on a one-degree-of-freedom vertical lift, which is in turn mounted on a Helpmate mobile base platform. A barcode scanner is mounted on the end of the arm to determine the closest box of books. This provides a basis for comparison between the desired and actual location of the robot. Currently this is the only global position feedback available to the robot. The composite system is coordinated by an on-board laptop PC. The PC sends task commands to, and receives task-level feedback from, the controllers for each of the subsystems.

Currently, the robot operates principally in an open loop mode whereby motor-level signals are given, and the robot moves as best as possible to the desired location. The current robot will need to be updated by replacing the laptop (which is now three years old) and the mobile base (because Helpmate is no longer an independent

business, and its technical support department can no longer provide the assistance, we will need to take the robot to the next level of performance). We will also need to incorporate a more sophisticated global feedback mechanism.

From Box to Scanner

Currently there is no mechanism for transferring books from the robotic retriever to the scanning station. This problem will be addressed in the proposed work. It consists of two distinct problems corresponding to the two modes in which the robot is intended to be used:

(a) Single Book per Box

This task will consist of opening the box containing a single book, and placing the book on the scanner.

(b) Multiple Books per Box

This task will consist of selecting a single book within an open box filled with bar-coded books, extracting the desired book, and placing the book on the scanner.

The scientific issues to be addressed in (a) are: (1) Appropriate design of book boxes and restraints imposed on books within the box so as to make them easier to manipulate than the original book itself; (2) The design, development, and implementation of fault-tolerant fixtures that will allow a box filled with a book to be gravity loaded into place on top of the page turner. The scientific issues in (b) are: (1) to localize the position of the book of interest within the box using bar code information; (b) to develop a force-sensing wrist/gripper that will pick a book out of the box and place it over the scanner without damaging it.

The Page Turner

There are page-turning devices but they are used in controlled settings, such as for patients undergoing rehabilitation (Danielsson and Holmberg 1994; Neveryd et al. 1995). Others require continuous adjustment to accommodate different types of paper (<http://www.4digitalbooks.com>). The technical issues in the design of the CAPM page-turner include the fact that some of the texts are quite old and fragile. We will investigate and build page-turning devices that will automatically accommodate a range of paper types from the materials at the Moravia Park facility of Johns Hopkins.

We will investigate three technologies for page turning. These technologies include: (1) electrostatic repulsion induced in the book by using a high voltage (but minute current) to cause the pages of the book to repel each other. The page-turner (which would also repel the book pages) will be a thin plastic strip with motors at one end. The strip will then sequentially insert between the separated pages and step through the book one page at a time. This method will require us to investigate dielectric properties of various kinds of paper; (2) A pneumatic suction-cup device will be investigated to determine how effectively it can grasp individual pages; (3) a rubber wheel actuated by a motor that turns over pages will be investigated. This is the simplest solution, though it may not be effective from the point of view of fragile books. Hence, a thorough comparison of methods and their relative merits will be performed. This will be done under various humidity and temperature conditions. The final system may comprise a combination of technologies, given the range of paper types that must be accommodated.

Senior Personnel Drewes, an expert in library preservation, will provide the relevant expertise to describe the range of paper types and will collaborate with co-PI Chirikjian to develop the page-turning system.

Proposed Schedule of Work for CAPM

In year 1, candidate designs for both box-to-scanner transfer scenarios will be investigated. A prototype page-turner will also be designed. Also during this year feedback mechanisms will be retrofitted on the current book-retrieving robot.

In year 2, the most promising designs box-to-scanner transfer mechanisms and page-turners will be implemented.

In year 3, the designs will be tested on books (and boxes of books) of different sizes, weights, and paper quality.

Proposed Development of GAMERA

With the existing testbed of digital humanities content, and the subsequent additions to this testbed through manual digitization and CAPM, it becomes possible to test and further develop GAMERA with a rich array of document analysis challenges.

Three general areas to be developed in this proposal are: creation of more plug-ins, integration with CAPM, and additions of various features to make GAMERA more usable and flexible.

More Plug-Ins (developed throughout duration of project)

It is important to realize that GAMERA's main advantage rests upon the concept of a framework or "shell" that allows domain experts to choose seamlessly from various document analysis methods. Thus, having a wide range of tools are desirable to allow the users to select appropriately for their tasks.

In addition to refining existing plug-ins, various new plug-ins will be implemented using existing published methods. These methods include: binarization, deskewing, layout analysis, feature extraction, character segmentation and classification. There is a wealth of published papers on the subjects proposed by researchers in document analysis and pattern recognition communities. For example:

- Sankur and Sezgin (2001) lists 44 binarization methods, categorized into six method groups: histogram shape-based, clustering-based, entropy-based, object attribute-based, spatial, and local.
- Cattoni, et al. (1998) reviews 23 papers grouped into following deskewing methods: projection profiles, Hough transform, connected-component (nearest neighbor) clustering, and correlation between lines, gradient analysis, Fourier spectrum, morphological transforms, and subspace line detection. Principal component analysis (Steinherz et al. 1999) has also been used for deskewing.
- Techniques used in document layout analysis include connected-components analysis, projection profiles, texture analysis, background analysis, smearing, morphology, and block classifications (Haralick 1994; Cattoni et al. 1998).
- Trier, et al. (1996) reviews nearly 100 papers on feature extraction methods including: template matching, graph description, projections, unitary image transforms, contour profiles, zoning, moments, spline curve approximation, and Fourier descriptors.
- Casey and Lecolinet (1996) reviews various character segmentation methods including projection analysis, connected component analysis, and applications of Hidden Markov Models.
- Classic classifiers (Duda et al. 2001) include nearest neighbors, statistical methods, decision trees, and neural nets. More recent techniques are support-vector machines (Vapnik 1995) and combination of classifiers (Xu et al. 1992; Kuncheva et al. 2001) including bagging (Breiman 1996) and boosting (Freund and Schapire 1995).

Many of these algorithms can be obtained as open-source software and can be seamlessly incorporated into the GAMERA framework as plug-ins.

Integration of CAPM

We will create a linkage between CAPM and GAMERA. Specifically, it is possible that the hardware system, CAPM, could be used to account for many pre-processing problems. For example, if a set of images has a systematic skew problem, it may be preferable to have the CAPM system adjust the scanning accordingly rather than rely upon deskewing methods only. Additionally, we intend to experiment with different resolutions for GAMERA to determine the "optimal" arrangement. That is, if 300dpi images are sufficient for document

analysis with GAMERA, then it is advantageous to use these (relatively) lower resolution images (e.g., reduced disk space needs). GAMERA will provide feedback to CAPM in this regard to facilitate optimal scanning of materials.

New Features

Features to be added to increase usability include:

- Component feedback and backtracking. (years 2-3)
- Add support for more image types (e.g. color and 3d). (years 4-5)
- Administration interface including managing documents, learning databases, and running GAMERA scripts on multiple computers with feedback in a centralized interface. This development, in particular, will allow GAMERA to scale for large-scale applications. (years 2-5)
- Facilitate transition from interactive processes to batch processes. (years 2-4)
- Training that would allow automatic component selection for optimization using component feedback and backtracking. (years 3-5)
- Development of domain specific interfaces given usability testing. (years 2-5)
- Add editor and other features to make GAMERA more of an interactive development environment. (years ICH?)
- Abstraction of syntactic and semantic analysis. (years 1-3)
- Develop or integrate document layout analysis tools (years ICH?)

Usability

The PITAC report on digital libraries correctly identifies usability as “a key issue in the design and operations” of digital libraries (PITAC 2001). The report lists several types of diversity that contribute to the challenge of designing usable digital libraries, including language and skill level. Different languages and skill levels result in users with different needs, tasks, and levels of expertise that necessitate customization in interface design. In addition, the diversity of digital library users is not only composed of different languages and skill levels, but also of different disciplines. A more comprehensive definition of “diverse users” would include academic discipline.

GAMERA’s users are diverse in this broader sense. The humanists and the librarians who use GAMERA have different training and needs. Of course, “humanists” comprise a diverse group of disciplines and technical skill levels. A scholar of ancient Greek might require a different features for an interface than an Art Historian. Furthermore, consider that the project team includes a technically trained humanist post-Doc from co-PI Crane’s Perseus Project and a post-Doc at Hopkins with more “traditional” humanist training. It is possible that the Perseus post-Doc may be more comfortable with a command line interface whereas the Hopkins post-Doc will require more GUI components.

As stated previously, GAMERA includes a goal to “leverage the user’s knowledge of the target documents to create custom applications rather than attempting to meet the needs of diverse users with a monolithic application.” In this same spirit, it is unrealistic to expect that a monolithic *interface* will meet the needs of diverse users. Usability testing and research throughout the development of GAMERA will foster the design and development of usable interfaces to accommodate a diversity of needs, skill levels, and disciplines. We have already considered methods and initiated plans to test the usability of the OMR interface with music researchers from the US and the UK. We will build upon these efforts to conduct extensive usability testing throughout the development of GAMERA.

Results from Prior NSF Work

Multiple members of the project team have received support from NSF for existing, related work. The proposed work is consistent with the research objectives and goals of the various project personnel.

DLI-2: Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music, Phase 2 (IIS-9817430)—Principal Investigator: Sayeed Choudhury, Johns Hopkins University, Co-Principal

Investigators: Ichiro Fujinaga, Tim DiLauro, Cynthia Requardt, Johns Hopkins University. The DLI-2 phase of the Levy Project focuses on the development of a workflow management system that will reduce the amount of human labor (and consequently cost) associated with large-scale digitization. Additionally, the project includes the creation of a suite of research tools to facilitate access to digital collections. The workflow management system will be demonstrated using a subset of the Levy Collection that has not yet been digitized.

The cornerstones of the workflow management system include the optical music recognition (OMR) system and an automated name authority control tool (ANAC). The OMR software generates a logical representation of the score for sound generation, music searching, and musicological research. This DLI-2 research has provided the foundation for GAMERA and supplied the evidence regarding the potential of GAMERA to address document analysis challenges. Since DLI-2 funding will end in April 2003, ITR funding beginning in October 2002 would provide continuity of research and development of GAMERA. Additionally, ITR funding would facilitate the inclusion of CAPM into the overall data capture framework.

DLI-2: A Digital Library for the Humanities (IIS-9817484)—Principal Investigator: Gregory Crane, Tufts University. The first two years of DLI-2 research have allowed co-PI Crane to build upon the Greco-Roman collections of the Perseus Project by establishing testbeds in six other areas of the humanities including: the history and topography of London; the archaeology of ancient Egypt; Shakespeare and early modern English literature; the American Civil War; the history of mechanics of antiquity through 18th century; and the Library of Congress' American Memory collections on California and the Upper Midwest.

These testbeds represent a wide diversity of cultural heritage materials that can be utilized to test and refine GAMERA. Additionally, the CAPM system will digitize other materials that co-PI Crane identifies for testbed development. There would be some overlap between this DLI-2 funding and the proposed ITR funding. However, this overlap will ensure continuity between efforts and will allow existing technically trained post-Docs from the Perseus Project to work with GAMERA.

Co-PI Chirikjian has received several NSF grants. His current NSF grant is **Diffusion Processes in Motion Planning and Control (RHA 0098382)**. Previous grants include:

Mathematics and Mechanical Engineering (DMS-9971696)—Principal Investigator: E.R. Scheinerman, Co-Principal Investigator: Greg Chirikjian.

A Paradigm for Inexpensive Service Robots (NSF-IRI/RHA 97-31720)—Principal Investigator: Greg Chirikjian

Design and Motion Planning of Discrete Robotic Systems (Presidential Faculty Fellow Award)—Principal Investigator: Greg Chirikjian

Design and Motion Planning of Discrete Robotic Systems (NSF National Young Investigator Award)—Principal Investigator: Greg Chirikjian

Finally, he plans to submit the following proposal to ITR program: **Massively Parallel Sensor-Based Approaches to Highly Articulated Robots and Haptic Interfaces—Principal Investigator: Greg Chirikjian.**

In the past, co-PI Chirikjian has investigated robot arms with discrete actuation and fixed topology as well as self-reconfigurable modular robots with topologies that can vary in a finite, but large, number of ways.

Due to the similarity between the kinematics of discretely-actuated manipulators and conformational issues in polymer science, the co-PI has been able to contribute there as well. The co-PI has developed a common framework based on Noncommutative Harmonic Analysis (Fourier Analysis on Groups) to address these problems. In the process, the PI has not only developed applications within robotics and polymer science, but

has also made contributions in the applied and computational mathematics literature. This in turn has led the PI to new robotics applications such as error propagation in linkages.

Prior NSF support also enabled the PI to publish extensions of his PhD thesis work in the area of “hyper-redundant” (snakelike) robotic manipulator inverse kinematics and motion planning as well as starting efforts in many new areas. The topic areas of the co-PI’s past work are:

- Discrete-State Robots and Spherical Stepper Motors
- Modular Self-Reconfigurable (Metamorphic) Robots
- Applications of Fourier Analysis on Groups
- Hyper-Redundant Manipulators and Mobile Robots
- Computer Aided Design and Solid Modeling

Although a number of the PI’s past contributions have been theoretical, many novel hardware designs have also come from his lab. The ITR funding for CAPM development represents another opportunity to design a novel hardware system (motivated by a unique goal).

Conclusions

Arguably, it is an ideal time to build this data capture framework given the open source movement, portable GUIs, powerful scripting languages and concept of generic programming. This approach will be essential for humanists dealing with the wide range of document analysis problems presented by cultural heritage materials.

The proposed work, which builds upon existing research of the project team, will result in a data capture framework that will start with a printed work and end with digital images, text, and musical content suitable for digital libraries. The resulting testbed, continuously augmented by CAPM, will offer a range of challenging document analysis problems to help maximize the functionality of GAMERA. Consequently, CAPM and GAMERA will comprise a data capture framework for cultural heritage materials that will accommodate a wide range of document analysis challenges.

It is worth noting again that the impact on the humanities may be the most significant. The humanities are a data intensive area of study: scholars complement close analysis of individual documents with searching and general document analysis. The field of classics, with a limited and heavily studied corpus, recognized the impact of automatic analysis a generation ago. The Theasaurus Linguae Graecae began developing a digital library of classical Greek texts in 1972. The results text corpus is now a foundational research tool – comparable in its significance to the much more widely known Genome database. The work proposed here would allow many other disciplines to expand their holdings. The humanities faces a generation of fundamental work as it creates for itself a new electronic scholarly infrastructure. The work outlined here attacks one of the key bottlenecks in such work.