3. Narrative

**Significance and assessment of need**

The majority of manuscripts accessible online today favors Psalters, books of hours, bestiaries, and other medieval manuscripts from Europe. Manuscript study in terms of Western texts is well-supported and enjoys high standards of critical scholarship. The Middle Eastern manuscript culture is less well known, primarily because the materials are far harder for qualified scholars to examine or study. Thus, by enhancing access, the possibilities for making a material change in our understanding of Middle Eastern cultures of the medieval, early modern, and modern times are great. Yale and SOAS seek to join collections, resources, staff, and expertise to create a united source of reference works related to Middle Eastern manuscripts and to develop the technical apparatus needed for connecting reference materials to original works. The proposed selection represents a wealth of intellectual interests that will highlight the contributions of Middle Eastern scholars, among them philosophers, poets, physicians, and scientists. The project partners plan to provide new and enhanced access to these important materials, as well as to develop a technical model that can be used by other libraries.

Increasing interest and attention has been given of late to the scholarly community that specializes in Arabic and Middle Eastern Studies. In the U.S., it is clear that there is an enormous rise in enrollments in Arabic classes. In May 2007, the Modern Language Association (MLA) published “Enrollments in Languages Other Than English in United States Institutions of Higher Education, Fall 2006,” in which Furman et al. presented comparative data on foreign language enrollments from 1998 to 2006. The authors reported that “Arabic continued its impressive expansion: from 1998 to 2002, it lifted its enrollments by 92.3%, and between 2002 and 2006 by a remarkable 126.5%.” In the UK, Ph.D. theses reflect the increase in enrollment and attention paid to Middle Eastern studies. In 1997, 55 theses were accepted; the average per year by 2006 was 86. Birmingham, SOAS, and Oxford are the top three institutions granting Ph.Ds. in Middle Eastern topics since 1997.

Concurrent to this visible growth, specific interest has been focused on the research needs of a scholarly community that was established centuries ago. In 2007, responding to an initiative in the United Kingdom regarding the importance of supporting Islamic Studies in higher education, the Joint Information Systems Committee (JISC) issued a call for an investigation into user needs within the field. The University of Exeter won the bid to complete the study, and in June of 2008 the project team published the results-based data extracted from online questionnaires, focus groups, and telephone interviews. In addition, the study’s authors reviewed reading lists from UK institutions, doctoral theses, and existing online gateways to Islamic Studies materials. The recommendation ranked first by the authors was the creation of a gateway to Islamic Resources, including primary texts, fully digitized Islamic manuscript catalogs, and reference tools such as dictionaries and Islamic websites.

A worldwide initiative has further concentrated on manuscripts in particular. Following the First Islamic Manuscript Conference held at King's College, Cambridge, in 2005, the conference participants encouraged the founding of a global association to coordinate the efforts of scholars and librarians working with Islamic manuscripts. In the following year, forty-five founding members established The

---


4 Auctherlonie et al., p. 9ff
Islamic Manuscript Association (TIMA), an international group pledged to protect Islamic manuscript collections and to support those individuals working with these collections. One of TIMA’s ongoing projects focuses on facilitating access to these manuscripts.  

**Related projects and other available resources**

A review of the digital landscape confirms that images of a far greater number of Western manuscripts are available. Additionally, it is clear that there is little connectivity between manuscripts and related reference materials. For example, the British Library (BL) offers online access to manuscript catalogs describing its collection of Western manuscripts, covering handwritten documents from pre-Christian, Classical, medieval and modern times. The BL site clarifies that the manuscript catalogs were scanned to create the online resource, but that it does not provide links to images of the manuscripts it describes. The patron is directed to use two BL websites to view or purchase images of manuscripts mentioned in the online catalog: *Digital Catalogue of Illuminated Manuscripts and Images Online.* Trinity College Cambridge and Yale’s Beinecke Rare Book Library (BRBL) have many Western manuscripts already available for online viewing. Trinity shows the first steps in linking catalogs to the original manuscripts. It provides online search functionality for M.R. James’ catalog entitled *The Western manuscripts in the library of Trinity College.* The catalog does not yet link to full sets of digital images; rather to sample images available via links in the catalog entries.

An examination of available digitized Arabic, Persian, or Islamic manuscripts leads to, among others, the British Library and the digitized collection from its spectacular Sacred exhibit in 2007. In another case, the digital library at Yale’s BRBL is growing at a steady pace and displays a small group of Arabic manuscripts. In addition, the BRBL digital library offers a choice of views of selected folios or, in other cases, such as Zamakhshari’s *Kitab al-maqamat,* a slideshow of the full set of images mimics the reading experience. At the University of California – Los Angeles, the Digital Collections section has a project underway to create finding aids for 1,506 Persian and Arabic manuscripts in its Minasian collection and to digitize approximately 304 manuscripts. The Melville J. Herskovits Library of African Studies at Northwestern University houses a collection of Arabic script materials from West Africa, containing over 5,000 items collected from Africa by Northwestern's professors. The website provides online access to a database catalog of the Herskovits holdings, which offers searching via metadata fields (e.g. title, author, subject) in English or Arabic. The Afghanistan Digital Library is a project developed by New York University which currently provides access to a collection of images from 362 books; the pilot project permits browsing and will soon include searching of the transliterated text. (For details on other related projects, see Appendix A.)

Other existing online resources provide descriptions of manuscripts held at leading institutions and offer directory information for these libraries, such as contact information and library hours. In the UK, the Archives Hub (http://www.archiveshub.ac.uk/index.html) provides a single point of access to descriptions of more than 20,000 archives held in 176 UK libraries, archives, or special collections departments. At present these are primarily at collection-level. The Archives Hub is one of the UK's National Archives'...
Network of related projects, such as ARCHON, the directory of national and international repositories which have substantial collections of manuscripts mentioned in the indexes to the National Register of Archives (http://www.nationalarchives.gov.uk/archon/). In the U.S., the Library of Congress hosts the National Union Catalog of Manuscript Collections (http://www.loc.gov/coll/nucmc/), which makes use of a search engine maintained by OCLC (http://www.loc.gov/coll/nucmc/oclcsearch.html). While helpful, the site requires a good deal of effort on the part of the patron, as well as an understanding of MARC21 metadata notation. For example, after the patron finds results to a query on manuscripts, she must then look at MARC21 metadata to locate the 040 field and its $a tag, in order to find the OCLC identifier of the library holding the manuscript. Armed with this code, the patron then navigates to the WorldCat Registry and its Advanced Search where she can enter the code in the OCLC identifier query box. Once submitted, the query returns information that, like the ARCHON display, shows directory information on the library that holds the desired manuscript.

Related digitization and finding aids projects bring the patron only so far. The present proposal intends to show proof of concept in creating a cross-referenced reading room experience in a virtual library by linking original works to related reference materials.

**Significance of selected materials**

The Yale-SOAS team proposes a pilot to create an archive of and a gateway to manuscripts. The selection for this pilot is composed of important manuscripts, related manuscript catalogs, and historical dictionaries held separately in the collections at Yale and SOAS. (Please see Appendix B for the full list.) In choosing the works, we have prioritized works of high intellectual value, in order to underscore the value of the materials in terms of benefits in areas of teaching and learning. We have reviewed together the MSS catalogs and dictionaries held by both libraries. We narrowed our selection by prioritizing 1) out-of-copyright resources, in order to avoid the time-consuming efforts needed for seeking copyright permission, 2) titles with missing volumes, so that we can join collections digitally as a result of this project, and 3) materials requiring preservation attention, for instance deacidification, boxing, or rebinding. The exception to this policy is Nemoy’s catalog, published in 1956, of MSS at Yale for which we have secured copyright permission in writing. (See Appendix C for the selection criteria.)

**Manuscripts**

Yale and SOAS have selected significant Arabic and Persian manuscripts. For example, Yale chose from its Medical-Historical collection. The library was founded in 1941 by the donations of the extensive collections of Harvey Cushing, John F. Fulton, and Arnold C. Klebs. The library owns about 25 Persian and Arabic manuscripts, mostly collected by Cushing. It is from this special collection that the project team has selected ten manuscripts of exceptional intellectual and historical value. As one example, Cushing MS Arabic 5 is a fundamental work on medicine entitled قانون في الطب (Canon of Medicine) from Ibn Sīna (Avicenna 980-1037). The manuscript contains 505 leaves with two leaves of notes in Arabic and Persian at the end. It is copied in a medium sized naskhi script. The leaves measure 21.5 x 31 cm (8.5 x 12.2 in) with a writing surface 14.7 x 22.5 cm (5.8 x 8.9 in). The text is gilt ruled, 39 lines to page, with catchwords and headings in red, blue and gold. Apart from his poetry and writings on philosophy, geometry, and astronomy, the Canon is Ibn Sīna’s best known work and is credited with disseminating medical theory and interpretation on symptoms and remedies among medieval physicians and scientists thanks to its translation into Latin by Gerardo de Cremona in the 12th century.

---

15 This catalog is the one exception to the team’s pre-1923 selection criterion. It was published in 1956 and lists many MSS already digitized and viewable in the BRBL digital library. Copyright permission for this catalog has been obtained and is on file at Yale University Library.

16 The full citation is: Kitāb al-Qānūn fī al-ṭibb / taʾlīf Abī ʻAlī al-Ḥusayn ibn ʻAbd Allāh al-Ḥasan ibn ʻAlī ibn Sīnā / كتاب القانون في الطب / تأليف أبي علي الحسين بن عبد الله الحسن بن علي ابن سينا.
From the SOAS collection, a representative selection is the Tadhkirat al-mulūk / Muḥammad Hāshim Āṣif (Rustam al-Hukamā) 1255 [1839], an historical work by the author of an important source for 17th-19th Century Iranian history known as Rustam al-tawarikh. A Manuscript judged by Arberry to be the original copy, it consists of 198 leaves [8 lines to the page] in a beautiful and clear Persian hand from Iran. In the absence of a published catalog of Persian MSS in SOAS, this work is known neither to Birgitt Hoffmann, who translated the Rustam al-tawarikh into German with a study of the author in Persische Geschichte 1694-1835 erlebt, erinnert und erfunden (Bamberg 1986), nor to Muhammad Mushiri, the author of the Persian published edition of the Rustam al-tawarikh, (Tehran 1973).

The combined selection represents a wealth of interests in an effort to highlight the intellectual contributions of Arabic and Persian scholars – philosophers, poets, physicians, and scientists – and provide new and enhanced access to an important scholarly community.

Manuscript catalogs

A manuscript catalog describes manuscripts held by an individual or an institution. Catalogs prepared by 19th and early 20th century scholars provide encyclopedic descriptions with a wide range of information rather than the customary inventory or lists found in today’s finding aids. They can describe the author’s life, cite particular passages of text in the original work, describe each manuscript’s provenance, physical composition and size, and locate it in an institution’s collection. Yale and SOAS have worked from a consolidated list of manuscript catalogs, describing Arabic and Persian manuscripts from renown international collections, to assemble an important selection for this pilot project.

Three Western catalogs in the selection represent joining transatlantic collections; both partners hold the catalogs but one or more volumes are missing from the shelves or need preservation attention. For example, the Bibliotheca Bodleiana codicum manuscriptorum orientalium is a highly desirable one because it records the manuscript holdings of the Bodleian library at Oxford, prepared from 1787 to 1835 by its editors John Uri (1726-1796), Alexander Nicoll (1793-1828), and E. B. Pusey (1800-1882). The work comprises two volumes, each measuring 45 x 28 cm (17.7 x 11 in). (As part of the Oxford University Library System, this catalog was included in the Google Book project. It is listed in Google Books with the notation ‘no view,’ indicating that either the images are not yet available or there is no full text equivalent of the work in the Google system.) Another chosen for the project is the 1894 catalog by Rieu, Supplement to the catalogue of the Arabic manuscripts in the British museum, important to scholars as it tracks holdings in the then-British museum now-British Library. One selected in order to fill in gaps missing from library shelves is Les manuscrits arabes de l’Escurial, décrits par Hartwig Derenbourg. More importantly, this catalog will provide a searchable listing of the Arabic holdings in El Escorial’s library, begun by Phillip II (1527-1598), who acquired his personal collection from the finest libraries and works within and beyond the borders of Spain. Catalogs of Arabic manuscripts assembled in the West predominate among those available for consultation. An exception is that compiled by Habīb al-Zayyāt and published in Egypt in 1902. Thus, within the Western catalogs, the selection includes text in Arabic, English, French, German, Latin, and Spanish.

Dictionaries

Yale and SOAS will each contribute three (3) historical dictionaries to the selection of reference materials. Why dictionaries? In order to create an integrated online resource for scholars working with manuscripts, dictionaries contemporary to the manuscript catalogs are essential. Often the vocabulary

---

17 The full citation includes أصف محمد هاشم. تذكرة الملوك / محمد هاشم رستم الحكمة
18 The full citation is Khazāʾin al-kutub fi Dimashq wa-ḍawāḥīhā : fi arbaʿatayzāʿ : Dimashq wa-Ṣaydnāyā wa-Maʿlūlā wa-Yabrūd /bi-qalam Ḥabīb al-Zayyāt
used in manuscript catalogs from the 19th century and earlier no longer appears in Modern Standard Arabic dictionaries available in print today. Further, there is a visible lack of authoritative online Arabic dictionaries, whether modern or otherwise. In some cases, canonical dictionaries can be found on the Internet but not in a usable or full text format, for example, Lane’s lexicon. An Arabic English Lexicon published in London between 1863 and 1893. A project to aid the study of the Qur’ān created in 2005 offers scanned images of Lane’s lexicon in downloadable PDFs – in single pages or by volume. Additionally, a Google Books search will bring the patron to a digital version scanned from Harvard’s collections in 2006. While both sites offer access to the important dictionary, neither digital avatar offers searchable Arabic text, only searchable full text in English. To compensate for the lack of searchable full text in Arabic, the Qur’ānic site has created a concordance by Arabic root called the Project Root List Online. In the case of Google Books, it is well understood that full text for non-Western script was not part of the project scope that included collections from Harvard University Library and other major collections of Middle Eastern materials. Today, however, it is possible to perform Optical Character Recognition (OCR) conversions on Arabic text and on print works with a mix of languages such as Lane’s lexicon, which includes English and Arabic.

**History, Scope and Duration**

**History**

Yale and SOAS boast distinguished histories in collecting Arabic and Middle Eastern Studies materials. Yale University Library holds approximately 12.5 million volumes, as well as information in numerous media, ranging from ancient papyri to early printed books to electronic databases. Yale was the first among American colleges and universities to encourage the study of Arabic and Middle Eastern literature and culture. When Edward Elbridge Salisbury, the first professor of Arabic, was appointed in 1841, he was the only scholar with this specialty in the United States. In the more than 150 years since Salisbury’s appointment, Yale has developed an extensive and internationally regarded collection of materials that includes 400,000 volumes supporting Arabic and Middle Eastern Studies. More than 3,000 manuscripts in Arabic, Persian, and Ottoman Turkish dealing with various subjects in science, Islamic law, and philosophy are housed in the renowned Beinecke Rare Book and Manuscript Library (BRBL) and the Medical/Historical collection at the Yale Medical School library.

Founded in 1916, the School of Oriental and African Studies (SOAS) is a college of the University of London, which was established by Royal Charter in 1836. SOAS is the only Higher Education institution in the United Kingdom that specializes in the study of Asia, Africa and the Near and Middle East. SOAS, with more than 4,400 students, 40% of whom are in postgraduate or research programs, maintains a position as the major UK center for studies concentrating on these world regions. Its library, designed by the architect Sir Denys Lasdun, opened in 1973. Recognized as a national resource, the library houses more than 1.2 million volumes, including significant archival holdings, special collections, and a growing network of electronic resources. The collection at SOAS, approximately 400 Arabic manuscripts and 400 Persian manuscripts, was cataloged before the World War II by A. J. Arberry, and the proofs of the catalog were shipped to Beirut in 1939. Unfortunately they did not survive the war, and the School has only Arberry’s notes and some galley proofs of the catalog. In 1980, Adam Gacek worked up the entries for the Arabic manuscripts with latter additions into a catalog which was published in a limited edition in 1981, and again in 1985 with corrections. IDC/Brill have made this catalog available online in PDF.

---

19 [http://www.studyquran.co.uk/LLhome.htm](http://www.studyquran.co.uk/LLhome.htm)
21 [http://www.studyquran.co.uk/PRLonline.htm](http://www.studyquran.co.uk/PRLonline.htm)
format. The Persian and Turkish manuscripts, which contain some unique and important works, have never been properly cataloged, and need to be made available to the scholarly community.

Yale and SOAS have collaborated to form a small but selective list of resources in Arabic and Western scripts from their collections. In addition, the project organizers have culled lists of existing digital copies of Arabic and Middle Eastern manuscripts held in repositories of prominent libraries around the world, for example in the digital library at the Beinecke as well as the British Library. We intend this selection to be a pilot project that is scalable and extensible to other collections on the partner campuses or in other libraries, such as the Digital Shikshapatri collection at Oxford and the Genizah manuscript collections at Cambridge.22

The proposed project builds on work done at Yale University Library in three digital initiatives related to the Middle East. First, the OACIS project (http://www.library.yale.edu/oacis/) laid the foundation by creating an electronic union list of serials published in and about the Middle East. From this beginning, the library team expanded the usability of the bibliographical catalog by digitizing full text articles from a selection of academic journals published in the Middle East. Second, with funding from the National Endowment for the Humanities (NEH), this digitization began with the Iraq ReCollection project, which converted nine (9) Iraqi journals (104,000 pages). (http://www.library.yale.edu/digIraq) Third, the full text entries from this digitization effort became the first deposited articles in a searchable repository developed as part of the AMEEL project. Like OACIS, Project AMEEL began with funding from the U.S. Department of Education. (http://www.library.yale.edu/ameel/) The AMEEL project also added a regional selection of academic journals spanning Tunisia to Saudi Arabia. Development work during the AMEEL project has produced the technical infrastructure for search, retrieval and display of the journal articles using the open source FEDORA repository software. At present, over 125,000 pages have been deposited with a goal of finishing 240,000 pages by the fall of 2009.23

While accomplishing these goals, the Yale project team has gained considerable expertise in Arabic text digitization. Starting in 2005 with the assistance form the digital staff at the Bibliotheca Alexandrina, the Yale team has since formulated its own digitization workflow.24 Further, the team has worked to automate as much of this workflow as possible, in order to keep labor costs low while producing high quality scanned images and OCR output. Additionally, the Yale team held a digitization workshop, geared toward the needs of U.S. academic libraries, at the November 2008 annual Middle Eastern Studies Association conference. (http://www.library.yale.edu/ameel/MESAworkshop/index.htm)

Digitization projects at SOAS are increasing in number and scope. The Endangered Languages Archive (ELAR), started in 2005 as part of an international network of digital endangered language archives, permits scholars to deposit documentations and descriptions of endangered languages. (http://elar.soas.ac.uk/) Begun in Fall 2008 with JISC funding, the Förer-Haimendorf project plans to digitize, research, catalogue and mount online approximately 20,000 photographs from the Förer-

22 For Oxford’s online access to Indian manuscripts, see http://www.shikshapatri.org.uk/. The Taylor-Schechter Genizah Collection at Cambridge holds 140,000 manuscripts fragments, mainly in Hebrew and Arabic. The University of Pennsylvania has joined Cambridge in this project. Further, the Special Collections division of The John Rylands University Library of Manchester holds a collection of around 11,000 Genizah manuscript fragments from the Ben Ezra synagogue in Old Cairo, written in Hebrew, Aramaic, Judeo-Arabic, Arabic and other Jewish languages.

23 For a look at the development history of the full-text repository including demonstration videos of search and retrieval, please see: http://www.library.yale.edu/ameel/project/ameel_demo.htm

24 The staff of the Digital Lab at the Bibliotheca Alexandrina (BA) is recognized for their expertise in Arabic text digitization, thanks to their ongoing participation in the Million Book project. The staff at the BA’s Digital Lab number approximately 120 specialists working during two shifts. In addition, the BA’s ISIS organization provides IT expertise and customized programming in three project teams to support all digitization efforts.
Haimendorf archive held at SOAS. (http://www.soas.ac.uk/furer-haimendorf/ ) Christoph von Fürer-Haimendorf (1909-1995) amassed an important collection of his own photographs, film, and written materials during fifty years of scholarship on tribal cultures in South Asia and the Himalayas. The project is part of a longer-term strategy to mount online the entire Fürer-Haimendorf archive, as well as other special collections at SOAS.

Since the library at SOAS has not previously worked on digitization projects involving manuscripts or printed text as other UK institutions such as Oxford and Cambridge, the Yale-SOAS partnership aims not only to address user needs by uniting essential reference material related to Arabic and Middle Eastern manuscripts, but also to share expertise gained at Yale and increase digitization capacities at SOAS, especially as the work relates to the digitization of Arabic text and the integration of digital resources.

**Scope**

In preparation for this proposal, we have gleaned information on the prospective users of the new archive from a comparison of users surveyed in the 2008 JISC report from Exeter along with those targeted in projects at British institutions such as Cambridge and Oxford, the British Library, and SOAS, as well as their U.S. counterparts at Yale, Michigan, and Harvard. The proposed selection and development are directed toward education and communities not necessarily versed in new technology that need improved access to research materials displayed in a usable and helpful fashion.

**Digitization endeavor**

A selective list of resources in Arabic and Western scripts from the collections at Yale and SOAS form the corpus for digitization. The team proposes to digitize sixteen (16) manuscripts held in the Medical/Historical collection at the Yale School of Medicine library and in the special collections at SOAS. These manuscripts comprise approximately 3000 leaves, ranging in height from 19 to 39 cm (7.5 to 15.3 in). The image capture for these rare and fragile materials from Yale can be carried out at the ITS-Medical Media services center without the need for transporting the manuscripts off campus. SOAS has a similar on-campus facility for the digitization of books and flat artwork housed in environmentally-controlled secure premises.

We also propose to digitize and convert to searchable full text fifteen (15) manuscript catalogs and six (6) historical dictionaries, representing approximately 16,800 pages in 40 volumes ranging in height from 24 to 45 cm (9.5 to 17.7 in) with text in Arabic, English, French, German, Latin, and Spanish. For proof of concept, the best manuscript catalogs for this project are those listing manuscripts already digitized or those selected for digitization as part of this project. For example, Nemoy’s catalog listing Arabic manuscripts in the Yale University Library cites at least twelve Arabic manuscripts already accessible via the BRBL digital library. (See Appendix D for a listing.) One particular entry from Nemoy is *al-Muttala’ / Abu Yahya Zakariya al-Ansari*. This manuscript is also cited in the Bodleian catalog (pars 2, Nicoll, ed.) and several references to other works by authors listed in Nemoy also appear. The Galen manuscript and the *Taqwīm al-abdān / rattabahu Yaḥyā ibn ʿIsā ibn ʿAlī ibn Jazlah* appear in both the Bodleian and the Rieu supplement. In particular, Rieu refers to the entry by J. Uri (Bodleian catalog, pars 1) regarding ibn Jazlah’s manuscript. Additionally, the Bodleian catalog, pars 2, includes an *Index librorum* where dictionaries by Freytag and al-Firuzbadi are listed. The catalog also provides an *Index vocabularum*, showing words of interest mentioned through the two-part catalog. This corpus of texts will help the integration team to focus on the technical challenges of mark-up, cross-reference linking, and citation creation.

**Integration project**

**Metadata generation**
For the manuscripts, we will encode metadata using TEI P5, the schema developed by the Text Encoding Initiative. This international standard adapts well to a mixture of formats, e.g. manuscripts, dictionaries, and catalogs. In preparation for this proposal, we consulted with metadata professionals (a Metadata Librarian and an Archivist specializing in Finding Aids). Both agreed that TEI was more suitable to the project than the Library of Congress standard, Encoded Archival Description (EAD), since we plan to create more than finding aids when working with the metadata. In addition, the content in the selected catalogs and dictionaries is more robust and narrative than that normally found in the lists and inventories common to finding aids. The team plans to make use of and modify as needed existing Open Source tools developed for TEI, such as ROMA, SACODEYL Annotator, and XML stylesheets distributed amongst the TEI community via their wiki.

Cross-reference links
The metadata generation will have a direct impact on the creation of active cross-reference links connecting the catalogs and dictionaries to the original works. Some of the challenges involved include: 1) conducting an analysis of the OCR-extracted text from the catalogs and dictionaries, 2) creating a usable index from the analysis, and 3) creating the programming needed to insert cross-reference links. The technical team for Project AMEEL will conduct the preliminary analyses using an Open Source product called AraMorph, which returns morphological tokens, or categorized blocks of text for lexical analysis. The resulting index will be reviewed manually, at first, to determine rules for extracting essential keywords, titles, place names, and subject headings. In addition, we will rely on language experts to develop crosswalks, or interpretive tables, to link modern spellings to the many transliteration schemas, varying over time, which are present in the selected materials. For example, the manuscript listings for ibn Jazlah and al-Suyuti in the Bodleian catalog and the Rieu supplement appear as Ali B. Djazla and Alsoiuthi. Further, in order to compile a full listing of existing digital copies of targeted manuscripts, we will employ two methods: 1) OAI harvesting of appropriately configured online databases, and 2) student workers, with language skills, to conduct Internet searches and review the OAI harvested results. The Yale team will share the methodology and results with the SOAS technical team in order to determine mutual processes for link creation.

Page viewing
As part of the AMEEL project at Yale, the technical team has developed a page turner to simulate the reading experience, with adjustable page size and navigation. Independently, the BRBL digital library team has crafted a slideshow approach for displaying manuscripts online. The technical team for the proposed project will work to adapt these methods, as well as other suitable Open Source methods, as needed to deliver a page viewer of high usability for library patrons. There are two basic outcomes from this effort: 1) to permit search word highlighting, and 2) to manage oversize displays from manuscript folios, catalogs, and dictionaries that allow the patron to work with more than one display at a time and well as enlarge specified sections of the works.

Durable URLs: book-bag functionality and citations
Increasingly faculty and students depend on access to research materials via the Internet for teaching and learning. This pilot project plans to supply methods for citing materials found in the proposed archive. We will provide durable URLs (Uniform Resource Locator) and offer citation links when displaying the materials. This will facilitate locating the materials at a later date, as well as citing the sources in published studies. In addition, we will offer a book-bag feature so that as a scholar works with the materials, she may create a snapshot of her research for downloading or temporary storage.

25 OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) Harvesters are software programs that search the Internet for metadata conforming to published OAI standards.
26 See : http://www.library.yale.edu/ameel/project/ameel_demo.htm
27 See a saved snap from the BRBL: http://images.library.yale.edu/digitalcollections/Groups.aspx?qg=202161379
Repositories and cross-collection searching

Yale’s Project AMEEL has developed a full-text repository using the Fedora open source framework. SOAS, like other UK academic institutions, uses EPrints for their digital libraries. The two technical teams will create the necessary tools to permit cross-collection transatlantic searching from the each partner’s archive and regardless of the entry point chosen by the library patron.

Duration

The Project Team seeks support for one year to show proof of concept in this pilot project, beginning in September 2009 and ending in August 2010. The work will be divided into two concurrent phases: Phase One will include all digitization tasks dealing with the printed reference materials and the manuscripts from the two library collections. Concurrently, Phase Two will focus on all integration activities, for example, the definition, tagging, and linking of metadata elements and cross-referenced materials, the expansion of the repositories to accommodate cross-collection searching, and the depositing and indexing of newly digitized materials. (More details are provided in Methodology and Standards.)

While we seek funding for this pilot project, both institutions are committed to the sustainability, beyond the term of this grant, of the system architecture and digital archive created during this proof of concept phase. This project makes use of lessons learned during AMEEL project at Yale University Library, such as its sustainability planning, including the creation of a budget for the four years following its grant term and the identification of fund raising opportunities with the assistance of Development Offices to seek funds from donors, federal agencies, and foundations.

Access to digital copies of Arabic and Persian manuscripts is limited. The coordination of related reference materials to digital copies of the original works is spotty. Very little experience in the digitization of texts printed in Arabic exists currently in academic institutions. Long-term sustainability of the project’s efforts directly responds to the large demand for creating digital resources in Arabic and Middle Eastern studies in the future, and will be achieved more efficiently by the lessons learned in the proposed proof of concept.

Methodology and Standards

Preparation and processing of materials:

Since we have completed a preliminary shelf check in order to compile a consolidated list of related holdings in both partner libraries, we will use this information in the first month of the project to pull all materials from the shelves, check them out from circulation if necessary, and deliver them to the respective digitization facilities. At the same time, we will update a tracking database, similar to the one used in Project AMEEL at Yale, to note accurate information for technical, descriptive, and administrative metadata tagging and to track workflow progress. Standards for image capture are listed in the following chart:

<table>
<thead>
<tr>
<th>media</th>
<th>resolution</th>
<th>ratio</th>
<th>archival format</th>
<th>Compression</th>
<th>number of digital copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>manuscripts</td>
<td>600ppi</td>
<td>1:1</td>
<td>TIFF</td>
<td>uncompressed or lossless compression; no LZW</td>
<td>4 (small, medium, large, thumbnail)</td>
</tr>
<tr>
<td>text—bitonal</td>
<td>300ppi*</td>
<td>1:1</td>
<td>TIFF**</td>
<td>CCIT Group 4+</td>
<td>1 ***</td>
</tr>
<tr>
<td>text—greyscale</td>
<td>300ppi*</td>
<td>1:1</td>
<td>TIFF**</td>
<td>LZW++</td>
<td>1 ***</td>
</tr>
</tbody>
</table>

* The resolution for text is set at 300ppi based on experience with text extraction using OCR software.
** Following the Project AMEEL model, the archival format will be TIFF; while the display format, i.e. from the repository to the page viewer, will be JPG to achieve faster online delivery.

*** The thumbnail will be generated at the time automated scripts deposit each digital file into the repository.

+ CCIT Group 4 is an image compression schema based on the "Comité Consultatif International Téléphonique et Télégraphique"), a telecommunications standard created in 1956

++ LZW is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch

We will use a combination of in-house and outsourcing – sharing lessons learned from previous projects – to complete all image capture. Varying costs will be factored into the budget to account for this hybrid approach. The manuscripts will be processed with appropriate supervision by related curators and Preservation librarians at on-campus facilities to limit the exposure of the fragile documents to outside conditions. The scanning of the dictionaries and catalogs may be outsourced, depending on their physical condition. Three OCR software products will be used: 1) Automatic Reader – OCR Gold from Sakhr Software Co., in Cairo, Egypt, 2) VERUS, the OCR product from NovoDynamics in Ann Arbor, Michigan, and 3) ABBY FineReader, an international company founded by David Yang for the automated translation of Russian dictionaries. ABBY FineReader is known for high accuracy conversion of text, especially in Western scripts. Sakhr and VERUS were developed for Arabic text specifically. VERUS, due to its original design, can handle a mix of languages and degraded documents better than Sakhr. On the other hand, Sakhr’s engine is based on a study of modern newspapers from the Middle East and thus recognizes a wider range of vocabulary. By incorporating two different OCR software packages into the digitization workflow, we can accommodate and manage varying conditions found in the selected materials. The OCR conversion of texts with a mixture of languages may require periodic modifications to existing workflows. The adjustments will be managed on a timely basis and will be documented so that workflow knowledge may be shared with other libraries.

We will follow established best practices for TEI mark-up as well as the Project AMEEL model to include Dublin Core for repository organization and MARCXML metadata for librarian perusal.28 Individuals with language expertise, as well as metadata training, will manage mark-up and quality assurance tasks. The Yale team will share best practices and guidelines with the SOAS team. For example, we will perform quality control checks on a statistical sample of all finished work, in accordance with American National Standards Institute ANSI/ASQ Z1.4-2003. A random sample equal to 10% of the total batch of files shall serve as the inspection sample for each file type. In tiered approaches employed in previous projects, a batch failing the 10% test is rechecked using a different sampling of 5% to determine final acceptance or rejection (and reprocessing) of the batch. Experience at SOAS suggests that, while post-digitization inspection is necessary, it is essential to build quality control into the digitization process in order to minimize the need to reprocess material. SOAS has adopted a strategy of embedding essential metadata within the image file using IPTC tags; this reduces the risk of orphan files (where image files cannot be found by the catalogue database or subsequently identified). It also facilitates the parallel development of the database (Phase 2) while digitization proceeds (Phase 1), with the embedded metadata being automatically ingested into the cataloguing database when ready.

** Organization of and access to materials

Metadata generation and Cross-reference links

All metadata created, for ingest to the Fedora archive, will be stored in XML format. We will begin with MARC21 records extracted from the partners’ Online Public Access Catalog (OPAC). Within the first month of the project, the technical team will convert these records into Dublin Core (DC) and MARCXML files at the title, volume, and author levels for manuscript catalogs and dictionaries; the manuscript records will have DC and MARCXML at the title, author, and accession level. The MARCXML file will be available when viewing searched materials, so that librarians may review encoded data. The customized Dublin Core (DC) file, added at the time of ingestion into the digital archive, will include tags to describe the resource as well as external and redirect instructions in the technical metadata for those materials with recognized cross-references. We have, initially, identified three Use Cases as follows:

♦ **Use Case #1:** a direct link exists from a digitized manuscript catalog to a digital copy of a manuscript freely accessible via the Internet.
♦ **Use Case #2:** a “See also” link from a digitized manuscript catalog to a digital copy of another manuscript from same author or, if possible, same subject.
♦ **Use Case #3:** a direct or a “See also” link from a digitized manuscript catalog to digitized dictionary entry.

We will begin the proof of concept with those Arabic or Persian manuscripts already available in Yale’s BRBL digital library. For those resources identified as existing in non-partner digital libraries, we will seek permission and technical specifications to create and maintain cross-reference links.

**Discovery: from search & retrieval to page viewing**

A single-design front door to the digital archive will be published on both partner web sites. Visual cues, in the form of icons, will be used to identify the location of the resources contained in the joined collections, or to point to resources held elsewhere. Title level metadata will also link from the digital archive to the partners’ OPAC. Searching, at a minimum, should be accomplished via a free-form text box in which the patron may enter a word or phrase. This search phrase will be analyzed against the TEI metadata as well as the full-text content. In addition, browsing functionality should identify for the patron what is available in the proposed manuscript gateway, i.e. by showing an expanding and collapsing tree-view of those manuscripts, catalogs, and dictionaries that are available. Retrieval of search results will include faceted classification, including the number of “hits” or returned results, as well as available and related keyword or subject headings. Search words in the display of full text from manuscript catalogs and dictionaries should be highlighted for easy viewing. Single-page or page turning with page resizing and navigation should be included to allow the patron to choose a preferred viewing mode and to increase reading comfort and for those with visual impairment.

**Cross-collection searching: EPrints / Fedora**

Project AMEEEL at Yale uses Fedora 2.2, an open source software product. The Fedora framework provides for OAI compatible harvesting for resource discovery, thus permitting other repositories to discover the newly generated metadata. EPrints, in use at SOAS and other UK academic institutions, is also open source software for OAI compliant repositories. Both repository approaches share commonalities, which will be explored to resolve the connectivity needed for searching the joined collections simultaneously. While the design of the front door to the proposed digital archive may appear the same, the underlying architecture at each library site will correspond to the requirements of the repository software in use. The technical team from both campuses will develop modules of code that can be adapted to both software approaches.

**Durable URLs and Citation creation**

---

29 For more technical information, please see: http://www.fedora.info and http://www.fedora-commons.org/.
30 For more on EPrints, originally developed at the University of Southampton School of Electronics and Computer Science, see http://www.eprints.org/.
Persistent identifiers are essential in developing a digital archive since storage media and formats are sure to change over the life of a digital library. Citation links will cease to function and cause frustration for library patrons without persistent identifiers. Some current standards for persistent identifiers include: 1) PURLs, or Persistent Uniform Resource Locators, which use an intermediate resolution service that points to the URL of the digital object; 2) DOIs: Digital Object Identifiers, are managed by an open membership consortium, which provides a name to identify a digital object that remains unchanged over the life of the object. The location of the object may change, but the name associated with it will not; and 3) Handles: a handle server is a naming management system that creates a unique identifier at the time an object is added to a repository. The Yale and SOAS technical teams will collaborate on the configuration of persistent identifiers compatible with both repository approaches.

Storage, maintenance and protection of data
The Fedora repository will reside on a Linux server at Yale; the EPrints repository will also be Linux-based but at SOAS. The archival TIFFs will be stored separately to the repository, but at each institution, which will arrange suitable off-site backup. Both teams will follow technical policies, including sensible naming conventions and file structure, so that content can be integrated effectively and efficiently as the project progresses. The servers will include systems management subscriptions to keep the server software protected and current. Information Technology support staff at the partner locations will regularly monitor server traffic, guard against attacks, and apply widely accepted security practices to development and maintenance servers. SOAS uses a mix of out-sourced and in-house server solutions, working in partnership with the University of London Computer Centre (ULCC). It is likely that ULCC will be used to host the EPrints repository while the archive of high resolution image files will be on SOAS’s own data storage solution.

Work Plan
Two segments of effort on parallel time lines will address digitization and integration goals. (Please see Appendix E for a chart of tasks and milestones.) Milestones Phase One: Digitize 16,800 pages in 40 volumes from 15 catalogs and 6 dictionaries, and 3000 leaves from 20 selected manuscripts. The two teams will meet within the first three months of the project to coordinate workflow, software, and hardware. The Yale team will organize and deliver a workshop for the SOAS group in image enhancement and OCR of Arabic and mixed language text. Milestones Phase Two: Using Yale’s servers, the Yale technical lead will establish, maintain, and archive a wiki for communication amongst the team members and for soliciting community comments, and a project web site. These will be accessible by both campuses and available by the end the first month to disseminate project information. The technical team will identify all metadata elements, define a project DTD for the TEI mark-up, and begin the analysis of extracted text. Two lines of effort will continue from November 2009 through the project end: 1) metadata generation activities including a Web-based tagging tool for use on both campuses, and 2) the archive presentation tasks including search and retrieval, page viewing, and citation generation. Final Deliverables: Scanning, processing, and OCR conversion tasks will be complete by the end of the project term. We will have established and tested links from the initial set of existing scanned manuscripts to newly digitized catalogs and dictionaries. We will have tested the OAI harvesting configuration so that the newly digitized materials can be discovered by other electronic resources. The project team will publish press releases to announce the accomplishments of the project and to promote the use of the new archive. The project team will publish its findings regarding the new digital collection and share project documentation via the project web site for use by other academic libraries.

Evaluation Plan: An evaluation of this project is important not only to meet the requirements of the funding agency for accountability of the expenditures of the grant, but also to share management information with US and UK libraries regarding the proof of concept, i.e. the virtual connection of original work to reference material. Since we are proposing a significant leap from today’s static finding
aids to an interactive research environment, the project evaluation will aim to gauge the following: 1) the success of the analysis of extracted text, and 2) the compilation of test results from simultaneous cross-collection searches conducted by a control group of students on both university campuses.

**Measures of Success:** 1) We have identified, in preparation for this proposal, twelve cross-reference links that should be created during OCR text extraction and consequent metadata creation. We will judge success when all twelve links are properly produced from automated scripts developed during the project. 2) Yale and SOAS will develop and conduct a usability study with control groups of undergraduate and graduate students. The study will focus on specific types of searches that retrieve data from each repository separately and the newly formed joint collection simultaneously. We will judge success when students can retrieve 80% or more of the search materials.

**Next Steps:** Since we are proposing a pilot to show proof of concept, upon successful completion of the project we would engage in planning the next appropriate steps for expanding from a pilot stage to a more robust project. Should there be new calls for joint UK-US collaborations, the Yale-SOAS team would seek to expand content by joining with willing partners in the UK, such as the Oriental Manuscripts and Rare Books collection at the Bodleian. If only US funding were available, we would begin with other libraries known for their Middle Eastern collections, such as the University of Michigan. On the technical front, next steps might include incorporating Web 2.0 features. This effort would involve, among other initiatives, fostering community building and adding user-generated content related to folksonomy tagging. 31

**Nature of Collaboration and Staff**
This core project team from Yale and SOAS represents a wealth of experience to benefit the proposed project. (Please see Appendix G for curriculum vitae of the principal members.)

**Principal Investigator:** Ann Okerson is Yale's Associate University Librarian with specific responsibility for Collections Development and International Programs. She is known nationally and worldwide for her work on the impact of electronic publishing, in the arena of costs, copyright, and licensing for the electronic environment. In addition to the PI role, Okerson will provide support for intellectual property and economic issues. (http://www.library.yale.edu/~okerson/alo.html)

**Project Director:** Elizabeth A. S. Beaudin is the Manager of International Digital Projects, Yale University. Beaudin received her Ph.D. from Yale University, writing her thesis on medieval love narratives from Muslim Spain. She is also a systems architect with 20 years of information technology experience. Beaudin, as OACIS technical lead, developed the software for its electronic union catalog. She serves as project manager and technical lead for Project AMEEL and Iraq ReCollection. Beaudin will direct all components of the project, monitoring progress toward meeting the specific project goals.

**SOAS Curator and SOAS Project Manager:** Peter Colvin BA, MA, DIP LIB, is Faculty Librarian for Languages and Cultures and Specialist Librarian for the Islamic Middle East at SOAS. Mr. Colvin first engaged with the Islamic World after his first degree when he spent from 1968-1970 in Kabul, Afghanistan, working as a UNA volunteer English language teacher. His second BA was in Arabic from SOAS and his MA was in Middle Eastern Area Studies at the same place. He has worked in SOAS since 1970, and from 1981 he has been the Specialist Librarian for the Islamic Middle East Section of the

---

31 Folksonomy is a term that covers collaborative tagging, resulting from categorizations, indexing, and other tagging tasks. The folksonomy terms are now increasingly displayed on a web page in a top-down presentation or in varying sizes, fonts, and styles to indicate relevance numbers whether in datasets resulting from a search or in articles covered in newspapers or blogs. By showing results from the full text content of the gazettes in this fashion, the patron can intuitively identify names and subjects that appear with more frequency in the content.
Library. For the last twenty years he has been the secretary of MELCOM-UK, the librarian association founded in the late 60s. In the 1970's he went on a book buying tour which took in all the countries of North Africa, the Levant, Iran and Turkey, and he spent a month in Tripoli, Libya, setting up a Library in the Castello. He catalogued and publicised a collection of early Arabic printed books which had been ignored in a corner of the Library. These books were published with indexes and an introduction as "Early printed books from Egypt at the Great Exhibition, London, 1851" [microform] / advisor, Peter Colvin. Leiden: IDC, 2002. He has also published 'Une histoire de la bibliothèque de la School of Oriental and African Studies.' La lettre d'information de la Bulac, 2. pp. 6-9., and 'Muhammad Ali Pasha, the Great Exhibition of 1851, and the School of Oriental and African Studies Library.' Libraries & culture: a journal of library history, 33 (3). pp. 249-259.

Yale Curator: Simon Samoeil has been Curator of the Near East Collection at Yale Library since 1990 and is a member of the Yale Council on ME Studies. Mr. Samoeil's specific contributions will include providing curatorial expertise. Mr. Samoeil has a broad and deep understanding of the field of ME Studies. Mr. Samoeil travels regularly in the ME and has developed numerous contacts there over the years. Before his arrival at Yale, Samoeil worked at the libraries of Harvard and the University of Pennsylvania and at King Fahd University in Dahran, Saudi Arabia.

SOAS Technical Director: Malcolm Raggett, MSc, DMS, has been IT Manager at SOAS and 15 years experience at managing and implementing IT systems including the SOAS digitization facility. Malcolm is currently technical advisor on the project to digitize the Furer-Heimendorf collection. Prior to this he has been an Information Services Director in a higher education college and has been in charge of audio visual services. Malcolm was a lecturer in further and higher education for nearly 20 years including teaching photography, which gives him a valuable insight to the academic value of visual resources and information as well as the technical requirements of the digitization process. He has been project manager for a number of small and medium sized projects, including the development of a digitization facility at SOAS, and has attended courses, delivered conference papers and led seminars on digital asset management. He has also been a passionate photographer for 35 years.

Yale Systems Programmer: Xinjian Guo has been a pivotal force in the technical development for the OACIS and AMEEL projects. Guo, a senior member of the central Library IT staff, brings over 15 years of systems experience to library-wide development projects. His principal duties on this project will include creating automated scripts for metadata generation and lexical analysis of extracted text.

SOAS Archivist: Susannah Rayner is Head of Archives & Special Collections at SOAS, and is a specialist in colonial and Commonwealth archives, having worked at the Oriental and India Office Collections (British Library), the Institute of Commonwealth Studies, and Rhodes House Library, Oxford. She has been involved in several digitization projects, including NOF-funded projects at Westminster City Archives and at Lambeth Palace Library, and the Getty-funded International Mission Photography Archive project. Her most recent project management experience was the creation of a new archives storage facilities (built to BS 5454:2000 standards), an expanded Special Collections Reading Room and new office accommodation at SOAS in 2007. She is currently involved in the establishment of a digitization suite at SOAS, and in devising a digitization program for the archives.

Yale Archivist: Bill Landis is Head of Arrangement, Description, & Metadata Coordinator. Bill has extensive digital library experience from his time at the California Digital Library. At Yale, as lead from the Manuscripts and Archives unit, he has recently managed the conversion of paper finding aids to a searchable Fedora archive.
Preservation and Collection Care Librarian: Ian Bogus is Head of the Collections Care unit in the Preservation department, the unit responsible for the physical well-being of reference and circulating collections at Yale. Bogus will bring Preservation expertise to the project as the selected materials circulate to and from the collection locations.

Dissemination
The teams will employ a variety of methods to disseminate information, solicit feedback, and encourage use of the proposed digital archive, such as:
1. A wiki for communication amongst the team members and to solicit community comments, a project web site to display the project proposal, project documentation for use by others, and regular announcements related to project activities.
2. Announcements and discussion of project work in publications that are:
   a. Scholarly:
      ♦ MELANET: A listserv devoted to Middle Eastern librarianship
      ♦ MELANotes: The journal of the Middle East Librarians Association
      ♦ MELCOM lis-middle-east: email list for MELCOM-International and MELCOM-UK
      ♦ MESA Bulletin: The Middle East Studies Association of North America’s journal of review
      ♦ MESA On-Line Newsletter: A quarterly publication on the MESA website
   b. Technical:
      ♦ DigLib: A discussion list for digital libraries researchers and librarians
      ♦ Fedora Commons: The discussion list for Fedora developers, project managers, and institutional management
      ♦ EPrints Community: The discussion forum for EPrints repositories
3. Laising with related professional associations such as The Islamic Manuscript Association (TIMA)
4. Talks on the project at MESA, MELCOM, and other professional meetings
5. Discovery by the curious scholar: Both teams will endeavor when publishing or delivering talks on the project to include details and/or instructions so that interested scholars may gain access and see the progress of the archive in development.