Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

Table of Contents
Abstract........................................................................................................................................... 1
Table of Contents............................................................................................................................ 2
1. Purpose........................................................................................................................................ 3
   Introduction – Gateway to Gazettes (G2G) Project................................................................. 3
   Concepts and processes............................................................................................................ 3
   Collection Building and Digitization ....................................................................................... 4
   Research and Development ...................................................................................................... 4
   Technical Knowledge Sharing ................................................................................................. 8
   Goals of Project G2G Related to Goals of TICFIA................................................................. 8
2. Extent of Need for the Project ................................................................................................. 10
   Recognition of National Need ................................................................................................. 10
   Increase in enrollments .......................................................................................................... 11
   Why digitize gazettes? ........................................................................................................... 12
   Parallel and Related Projects ................................................................................................. 13
   Specific Content to be digitized ............................................................................................... 16
3. Significance............................................................................................................................... 19
   National and International objectives .................................................................................... 19
   Theory, Knowledge, and Practices in related areas of study .................................................. 20
   Dissemination of project materials ......................................................................................... 22
4. Project Design........................................................................................................................... 24
   Milestones ............................................................................................................................... 25
   Final Deliverables: .................................................................................................................. 27
5. Project Personnel ...................................................................................................................... 27
6. Project Services........................................................................................................................... 30
   Material Processing.................................................................................................................. 32
   Shelf Checking and Workflow................................................................................................. 32
   Standards................................................................................................................................. 33
   Optical Character Recognition (OCR)..................................................................................... 34
   Quality Assurance.................................................................................................................... 35
   Organization and Access........................................................................................................... 36
7. Resources .................................................................................................................................. 38
8. Project Management .................................................................................................................. 39
   Commitment of the project team ............................................................................................. 39
   Sustainability............................................................................................................................ 40
9. Evaluation .................................................................................................................................. 40
10. Budget Narrative....................................................................................................................... 43
11. Appendices............................................................................................................................... 46
1. Purpose

*Introduction – Gateway to Gazettes (G2G) Project*

Yale University Library requests funding to build a digital collection of Middle Eastern official gazettes and make this resource available online in full text to researchers (notably in Law, History, Social Science, and Middle Eastern Studies) and to extend access to the public, commercial sector and law firms, GOs and NGOs. Official gazettes are state publications promulgating laws and regulations. They contain bills, acts, statutory instruments, regulations, decrees, orders, resolutions, international treaties, and selected court reports. In addition, all gazettes include formal government notices. Foreign legislative collections are important to legal research and are also relevant to a wide range of research in the humanities and social sciences. Research areas such as Middle East politics, energy, migration, trade, investment, human rights and development increasingly require reference to foreign law. To meet these needs effectively, it is essential for U.S. researchers to have access to extensive runs of serial primary sources (especially official gazettes) for authoritative research, available in one central place.

*Concepts and processes*

Previous Arabic digitization work at Yale has included a project to digitize the full text of nine Iraqi journals under the “Recovering Iraq’s Past” initiative of the National Endowment for the Humanities (NEH). In turn, the "Iraq ReCollection" project forms part of a more ambitious ongoing strategic ME effort at Yale, which focuses on the digitization of Arabic text in Project AMEEL (Arabic and Middle Eastern Electronic Library). AMEEL is also beginning to provide consolidated access to resources of academic value via the Internet by indexing and searching
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

full text content of print materials from the ME.\(^1\) OACIS, an electronic union catalog of serials from or about the Middle East, funded by the Department of Education, laid the groundwork for Yale’s work on Arabic text digitization.\(^2\) This proposal hopes to build on these initiatives by creating a Gateway to Gazettes Project (G2G). (Please see Appendix A for a glossary of the acronyms, initiatives, and terms used throughout this proposal.)

**Collection Building and Digitization**

The project vision is to establish a foundation for sharing distributed digitized collections of ME official gazettes across national boundaries, thereby acting as a model for other collections of foreign legislation and laws. Using Yale’s digital repository structure and standards will allow for flexibility and future expansion. Future participants will be able to make use of this infrastructure and practices to add their digitized collections of ME official gazettes as and when funding becomes available for digitization. As currently published official gazettes of some countries are becoming available online, the transition to electronic access to these holdings is becoming essential. The future dream for the entire project is a complete collection of ME official gazettes, such as the historic print holdings of the 18 ME countries covered in the US, the UK, and in other countries.

**Research and Development**

Official gazettes often lack consistent indexing, appear at irregular intervals, and vary in size over the publication history. For example, in the case of the Syrian Gazette, each volume includes a number of issues, typically with several different sets of pagination (e.g., a separately-

---

\(^1\) The Iraq ReCollection project, funded by the NEH from 2005-2007, focused on the digitization of nine Iraqi journals in the Humanities. Project AMEEL, begun with funds from the Title VI TICFIA program in the Department of Education, seeks to create a central entry point to scholarly materials on the Middle East by digitizing full text content from selected serials, encouraging Document Delivery exchanges among ME academic libraries, and integrating existing digital resources by means of a Fedora framework repository.

\(^2\) Online Access to Consolidated Information on Serials (http://www.library.yale.edu/oacis/)

- 4 -
numbered table of contents for one or more sets of pages, often intermingled with one another, some of which continue pagination sequences from earlier issues and some of which begin new pagination sequences). Because of the nature of the gazettes, the university team will analyze the materials and define an organizational structure, or metadata content model, suitable not only for use with both selected gazettes, but also scalable for the future. The Fedora infrastructure will be used for search, retrieval, and display. (The Fedora framework is explained further in the Project Services section below.)

To do this, we will create metadata and finding aids that are consistent in structure and robust in descriptive information. The research and development components of the present proposal will first construct a web-based tool to assist the Yale team in creating multi-lingual finding aids for its own collection and for deposit into a searchable architecture based on the metadata content model mentioned in previous paragraph. Official gazettes are often poorly indexed or not indexed at all. Thus, the creation of a metadata/finding aids tool will assist the Yale Library, as well as other libraries in the future, to organize and index the holdings more fully.

The same finding aids will later assist scholars in gaining access to the primary resources. While full text digitization permits searching in the language of origin, it does not allow for searching in related languages. Rather than spend time on machine translation of the full text content, we propose the creation of multi-lingual descriptive metadata (addressing Arabic-English first, adding French later) for use in patrons’ finding aids so that the new student entering Middle Eastern studies, lacking sophisticated command of Arabic, can locate resources as easily as a seasoned faculty member.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

In terms of research, we have set the following two goals. First, we hope to extract tokens, place names, and alternate definitions via AraMorph from OCR-extracted text to construct descriptive metadata and add this with the digital object, as the digitized gazette issue is deposited in the repository.\(^3\) We plan to design and develop a web-based tool for team members to initiate this metadata generation process. AraMorph is an Open Source tool currently used as a plug-in to the Lucene search engine in the Fedora repository for the AMEEL project.\(^4\) Currently, AraMorph provides indexing data to the search engine so that the AMEEL search interface can accept and process search terms against Arabic full text extracted from scanned images. The AraMorph java applet, however, does much more. It conducts morphological analyses of text and returns a variety of data that can be exploited for organizational purposes. (See example of extracted tokens and alternate definitions in Appendix B). 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.

Additionally, we will make use of new technical approaches to the display of metadata, such as folksonomies seen in Web 2.0 development. 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

---

\(^3\) For example, if the search string is كتاب for book, AraMorph returns morphological tokens, or categorized blocks of text for lexical analysis. From this analysis, related or alternate definitions, such as author, writer, and village school, can be extracted.

Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

In this fashion, the patron can intuitively identify names and subjects that appear with more frequency in the content.  

Second, to enhance the potential for cross-repository searching, we will create multilingual finding aids (Arabic-English, adding French later) as the descriptive metadata records are prepared for the repository. One way of accomplishing this goal is to harvest metadata records from one repository and deposit the harvested dataset into another using OAI. Often, though, OAI harvesting methods return datasets that are large and that contain records not directly applicable to the subject matter of a specialized project. We plan to create the finding aids and metadata when new or modified digital objects are deposited into the repository.

In order to integrate the newly generated metadata with existing digital copies of targeted word groups, we will employ two methods: 1) OAI-PMH harvesting of appropriately configured online databases, and 2) student workers, with language skills, to conduct Internet searches and review the OAI harvested results. We will begin the OAI discovery with a specific topic source, the Avalon Project at Yale Law School. (More detail is provided in the Parallel and Related Projects section ahead.)

Additionally, we will automate the exchange of these metadata files and develop a tool for the integration of the records into the repository at Yale. For the long term, we propose to develop these tools so that participating libraries in the future can deposit their digitized collections of ME official gazettes into their digital repositories and share descriptive metadata, based on full text content, with all other participating repositories.

---

5 Folksonomy is a term that covers collaborative tagging, resulting from categorizations, indexing, and other tagging tasks.
6 OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) Harvesters are software programs that search the Internet for metadata conforming to published OAI standards. For more information on OAI harvesting, please see [http://www.openarchives.org/](http://www.openarchives.org/).
Technical Knowledge Sharing

Very little experience in the digitization of texts printed in Arabic exists currently in academic institutions. The large demand for creating digital resources in Arabic and ME studies today and in years to come will be able to be achieved more efficiently by the lessons learned and the tools created in the proposed project. Rather than creating standalone, non-standard resources, the project team will conduct research, develop, test, and implement tools that can be shared and modeled for future collaboration. Following on Yale's Project AMEEL initiative, therefore, an important task for the proposed project will be to compile and publish documentation on the successful practices used throughout the project. In the above respects, the proposed project acts as a pilot to solve the challenges presented by the gazette format and to create tools usable by the libraries and other future participants.

Goals of Project G2G Related to Goals of TICFIA

The TICFIA legislation (section 606 of Title VI, part A, of the Higher Education Act of 1965) awards grants to "institutions of higher education, public or nonprofit private libraries, or consortia of such institutions or libraries, to develop innovative techniques or programs using new electronic technologies to collect, organize, and preserve, and widely disseminate information on world regions and countries other than the United States, that address our Nation's teaching and research needs in international education and foreign languages." Project G2G will achieve numerous goals, not limited to those sought by the Title VI/Section 606 grant requirements. The goals identified in the TICFIA web site and the relationship to the proposed activities follow:

**Goal 1: To facilitate access to or preserving foreign information resources in print or electronic forms:**

No truly comparable project for making these gazettes widely available exists. Related work has focused on print editions only, many of which are in fragile condition, or on microform
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

copies. Past and present projects in the US and the UK have attempted to document the existence of foreign legislation in official gazettes. Primarily these projects have mapped locations of gazette holdings by country of publication.

Goal 2: To develop new means of immediate, full-text document delivery for information and scholarship from abroad:

A search via Google Books shows that neither the Syrian Official Gazette nor the Palestine Gazette have been digitized by any Google- (or other large-scale digitizing) sponsored initiative. Additionally, neither Google nor any other large-scale digitizing project has to date produced full-text content searchable in Arabic. The Palestine Authority has begun to digitize modern issues of the Palestine Gazette for publication on its web site. No searching apparatus exists for these recent issues. In addition, the web site for the Ministry of Justice has undergone significant redesign such that materials accessed in 2008 are no longer available. Other related projects do not include full text searching or online access to the collections. The Avalon Project at Yale Law School, which contains much legal material of the period and supplements these gazettes, is an important exception; in Avalon, online searching is available. We propose to link to the Avalon legal documentation collection.

Goal 3: To develop new means of shared electronic access to international data:

The proposed collaboration with the Yale Law School will make available important and relevant materials held across the Yale University system. By making these materials available online, the Yale Library project will enhance access to important international data related to law making and nation building.

Goal 4: To support collaborative projects of indexing, cataloging, and other means of bibliographic access for scholars to important research materials published or distributed outside the United States:

Both the Syrian Official Gazette and the Palestine Gazette were printed and published outside the U.S. In many cases, the contents of these two collections of legislative documentation are either unavailable or exist in only the very elite universities – mostly in fragmented holdings. They are costly to acquire in terms of staff time, due to the complexities of governmental publishing offices, and they present a challenge for users, as the gazettes are either poorly indexed or not indexed at all. Title changes have occurred in both gazettes over the course of their publication history.

Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

**Goal 5:** To develop methods for the wide dissemination of resources written in non-Roman language alphabets:

The gazettes are primary sources from Arab law makers that can provide deep background information on laws, regulations, and other legislative activities. Both selected gazettes are printed in a mixture of languages, either Arabic-English or Arabic-French.

**Goal 6:** To assist teachers of less commonly taught languages in acquiring, via electronic and other means, materials suitable for classroom use:

An increase in access to valuable reference materials will respond to the growing demand for materials in the vernacular as enrollments in Arabic courses, especially at the undergraduate and graduate level, continue to rise.

**Goal 7:** To promote collaborative technology based projects in foreign languages, area studies, and international studies among grant recipients under this title.

An added value of this digitization project involves working with microform as well as printed text. At Yale, we have acquired experience in digitizing printed Arabic text along with OCR conversion of text to searchable content. We will apply these lessons learned to microform in order to complete a crucial period in the history of two nations and to share technological experience by inviting similar library teams for a seminar in Year Three.

**2. Extent of Need for the Project**

The project team believe that this effort responds to a need for consolidated access to primary materials from the Middle East that shed light on its past at a crucial time of nation building.

**Recognition of National Need**

The Secretary of Education, as required by the Higher Education Act (HEA) consults with the heads of federal agencies to receive recommendations regarding areas of national need for expertise in foreign languages and world regions. Current documentation provided on the TICFIA web site lists these recommendations. The following federal agencies highlight Arabic as a foreign language and the Middle East as a world region:

| Department of Agriculture | Arabic, 2nd among 8 languages listed as “most vital to our country’s future” |

8 [http://www.ed.gov/about/offices/list/ope/iegps/consultation.pdf](http://www.ed.gov/about/offices/list/ope/iegps/consultation.pdf)
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

<table>
<thead>
<tr>
<th>Department of Commerce</th>
<th>Arabic, 1st among 5 languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department of Defense</td>
<td>Arabic</td>
</tr>
<tr>
<td>Department of Health and Human Services</td>
<td>Arabic, 3rd of 7 languages</td>
</tr>
<tr>
<td>Department of Interior</td>
<td>Arabic, 1st of 7 languages, and highlights North Africa and the Middle East</td>
</tr>
<tr>
<td>Department of Labor</td>
<td>Arabic, 1st of 6 languages; Middle East, 4th of 6 world regions</td>
</tr>
<tr>
<td>Department of State</td>
<td>Arabic (Modern Standard, Egyptian and Iraqi) as 1st of 6 “super critical languages”; Arabic (other than MSA, Egyptian and Iraqi) 1st of 15 “critical” languages</td>
</tr>
</tbody>
</table>

**Increase in enrolments**

In the last decade, it has been clear that there is an enormous increase in enrollments in Arabic classes in the U.S. In May 2007, The Modern Language Association 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%”. Their data showed growth in both undergraduate and graduate courses as can be seen in these figures extracted from Table 2a.

<table>
<thead>
<tr>
<th></th>
<th>Undergraduates</th>
<th>Graduates</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>3,902</td>
<td>8,194</td>
<td>18,650</td>
</tr>
</tbody>
</table>

The figures support the fact that the study of Arabic continues to grow and, by extension, so does the need to provide access to reference materials used in the teaching of the language and in the investigation of its intellectual contributions to the world of knowledge and the foundations of nation building. The U.S. Department of Education recognizes the need for supporting these languages in U.S. university programs, notably through National Resource Centers (NRCs) and Foreign Language and Area Studies (FLAS) fellowships. These two

---


10. Furman et al., Table 2a, p.15.

11. [http://www.ed.gov/about/offices/list/ope/iegps/lctlanguages.html](http://www.ed.gov/about/offices/list/ope/iegps/lctlanguages.html)
Proposals to TICFIA: Yale University’s Gateway to Gazettes (G2G)

initiatives are part of the Title VI programs, celebrating its 50th anniversary in 2009. The present grant proposal plans to provide access to official gazettes to assist students supported by such Title VI strategies.

Why digitize gazettes?

Official Gazettes offer a wealth of information to researchers in the humanities and other related fields. Their digitization is significant and warranted for two main reasons: one relates to the nature of the material itself and the other to its content.

The nature of the material:

- The gazettes are costly to preserve, as they are printed on low quality paper and sometimes arrive damaged. Complete library holdings are thus uncommon.
- They present a challenge for users, as they are either poorly indexed or not indexed at all. Title changes have occurred in both gazettes. For example, the Syrian Gazette began as “Al-‘Asimah” and changed to “Al-Jaridah al-rasmiyah”. The Palestine gazette’s title changed from “Official Gazette, O.E.T.A. (South)” to “Official gazette of the Government of Palestine” then to “the Palestine gazette”. Various supplementary volumes were issued at irregular points in time.
- They are costly to acquire in terms of staff time, due to the complexities of governmental publishing offices.
- They are complex to manage and costly to house, in that many are published daily and in variable sizes.
- They are problematic to arrange on the shelf, as they are issued in numerous series and sub-series.
- It is problematic to maintain the catalogue records, as the gazettes not infrequently change title.

The Gazettes’ content

- The gazettes are an authoritative statement of the law in case of any dispute or inconsistencies.
- They are the preferred source of correct legal citation.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

- They represent the most up-to-date (and for some jurisdictions the only) published source of primary law.
- They are a key source of information about the status of legislation, since publication in the official gazette is for many jurisdictions the act that causes a law to come into force, thus giving researchers in history, law, and area and cultural studies, social sciences, and other fields, a unique look at the evolution of governments.
- They are a unique source of historical information for researchers in countries where the gazette is the only source of primary law.

Parallel and Related Projects

No truly comparable project for making these gazettes widely available exists: parallel work has focused on the print edition only. Past and present projects in the US and the UK have attempted to document existence of foreign legislation (official gazettes). These projects are based on either acquisition programs or mapping of locations and holdings, which are mostly shown at a coarse level of detail.

1. The foreign official gazettes program of the Center for Research Libraries (CRL)12 in the US aims to collect and preserve official gazettes worldwide up to 1995. However, this program does not attempt to provide full text searching or online access to the collections. Our Middle Eastern official gazettes project will offer online access to the collections’ full text.

2. Jules Winterton, Associate Director and Librarian of the Institute of Advanced Legal Studies, who aimed to make a comprehensive collection of foreign official gazettes available in the United Kingdom to support research, proposed a distributed national collection of printed foreign gazettes. The FLARE13 group welcomed the proposal and agreed to create and maintain a union list of official gazettes on its website, which mapped the location of holdings nationally. This list covers holdings of European official gazettes, mainly from Central and Eastern Europe but does not include ME official gazettes.

12 The Center for Research Libraries (CRL) is a consortium of North American universities, colleges and independent research libraries (http://www.crl.edu/log/index.asp). The project information is nearly 2 years old and, while it represents a singular effort, the data are incomplete.

13 FLARE: Foreign Law Research, is a collaboration between the major libraries collecting law in the United Kingdom (http://ials.sas.ac.uk/flare/flare.htm).
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

3. A FLAG\textsuperscript{14} project, which is a collaborative collection management project for foreign, comparative and international legal material, has also improved access to printed gazettes within UK libraries, including ME gazettes.

4. Listings of resources about Middle Eastern legislation exist online. For example, beginning at Cornell’s Middle East & Islamic Studies Collection on the web page for Laws, Legal Systems, the patron can go to Cornell’s Law School site where another listing called ‘World Legal Materials from the Middle East’ provides links by countries.\textsuperscript{15} These links take the user to specific agencies within the countries’ governments. A regular problem, though, continues to be access to some of the listed sites. The link to the Palestinian National Authority (http://www.pna.gov.ps/) produces a “Directory Listing Denied” error, meaning that access has been denied.

5. Other integrated resources online would link well to the primary sources from the selected gazettes. For example, Columbia’s Middle East Studies Internet Resources organizes electronic bibliographic resources and research materials much like the metadata directory in Yale’s AMEEL project.\textsuperscript{16}

   Our project proposes a newer and more comprehensive approach to official gazettes. The scope of the project concentrates on official gazettes from a specific geographic area. Additionally, the scope will be expandable to include legislative resources beyond the US. Ease of access to significant ME resources, appropriate long-term preservation/storage, and the nature of the material are driving motivations for the proposed digitization project.

   There is immense potential for further digitizing and interoperability/cross-searchability of collections. Very limited (incomplete) digitization of gazettes has occurred at Harvard University and the University of Michigan; however these collections require campus credentials

\textsuperscript{14} FLAG: Foreign Law Guide is a collaborative Internet gateway to the holdings of foreign, international and comparative law in UK universities and national libraries ( http://ials.sas.ac.uk/library/flag/flag.htm ).

\textsuperscript{15} http://www.library.cornell.edu/colldev/mideast/law.htm

\textsuperscript{16} http://www.columbia.edu/cu/iweb/indiv/mideast/cuvlm/index.html
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

for access. The official gazettes collections in the UK are much more complete: covering gazettes of 18 ME countries, including the historic Persian Gulf gazettes of the British administration in that region. (See Appendix C for the UK landscape.) These holdings are scattered among 11 specialized, academic and national UK libraries, which, along with U.S. libraries, could potentially over time join this digitization project. And, as some current official gazettes are becoming available online, this would be an appropriate time to look at the historic print holdings and to make a full transition to electronic access to these holdings. Thus, Yale Library will conduct the project with the long term possibilities of future expansion in mind, either with U.S., UK or other international partners.

At the same time, we believe that OAI harvesting methods can serve to link these newly digitized materials with existing resources. We propose to create OAI-discoverable finding aids, but also aim to use OAI as part of the scalable model we will develop; in this case, to create links from the newly digitized collection of gazettes to other related digital archives. For example, the Avalon Project: Documents in Law, History and Diplomacy, a project begun in 1996 at Yale Law School, provides digital copies of documents relevant to the fields of Law, History, Economics, Politics, Diplomacy and Government. The Avalon collection spans centuries of intellectual contribution to these fields, from Roman edicts and the code of Hamurabi up to government documents of the 21st century. Some helpful entries include a copy of the Palestine Mandate as well as British publications into the 1950s through to documents such as The Mitchell Plan from the Sharm al-Sheikh discussions in 2001. Dealings with Syria figure throughout 20th and 21st century digital documents, covering relations

---

Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

with Israel, State Department findings, and United Nations Security Council Resolutions. (See **Appendix D** for a sampling of documents held in the Avalon Project.)

By digitizing the holdings of key ME official gazettes in the US, we are appealing to numerous other institutions around the world (particularly in the UK and Europe), where runs of gazette resources exist, ultimately making these distributed collections accessible from one central point: a web gateway. Union catalogs focusing on gazettes show that holdings at other U.S. institutions are scattered or incomplete. For example, the CRL catalog of gazettes shows that both Harvard and the University of Michigan have holdings of the Syrian Gazette. Far fewer and less complete holdings for the Palestine Gazette (British Mandate) are available in the U.S. Yale Library holds the complete runs of the selected gazettes.

**Specific Content to be digitized**

The Yale University Library holds consecutive yearly volumes of the Syrian Gazette – the only such complete run in US libraries – from the first days of its activity in 1919 up to the present. Content to be digitized in this proposal includes: 1) the Syrian Gazette (from the French Mandate), starting with the first volume published in 1919 and continuing through to 1948, a pivotal year in ME history and held in full in the Yale Library; and 2) the Palestine Gazette (British Mandate), which is held by Yale in 40 reels of microfilm and several supplemental print volumes from 1919-1948. Detailed checking, down to the page level, will ensure full runs of both titles, even if missing pieces have to be cobbled together from other holding libraries.

Using the AMEEL experience as a foundation for this project, Yale will digitize and perform Optical Character Recognition (OCR) conversion on 59 oversize volumes plus six supplemental volumes of the Syrian Gazette, starting with the first volume published in 1919 and
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

continuing through to 1948. The approximate page count for this pilot phase is 70,000 pages. Most of the digitization will be carried out by a vendor selected for excellence in digitization services, chosen from those who are experienced with the Arabic language. We have also factored in time estimates to accomplish manual digitization in house should this be necessary based on binding or paper conditions. (More details are provided in the Project Plan section.)

A detailed shelf check of 57 volumes from The Syrian Gazette has already been completed at Yale. (Please see Appendix E for the report.) The volumes, in the majority, are oversized and are printed in Arabic and French. The dimensions for those volumes surveyed vary from minimum sizes of 9.63 inches (244.6mm) tall by 6.63 inches (168mm) wide to the maximum of 13.25 inches (336.5mm) wide by 9.38 inches (238mm) wide. Earlier volumes of the Syrian Gazette (1919-1939) may present older binding styles, fragile paper or other conditions that may require manual scanning. We plan to take advantage of robotic scanning for the majority of the selected volumes to enhance image quality and subsequent OCR accuracy as well as to keep scanning costs low.

The Yale Near East curator has worked with the government on other information initiatives and has secured permissions through known sources for other projects. (Please see Appendix F for a sample permissions letter used in the AMEEL project at Yale.) Regarding the Syrian official gazette, the Yale Near East curator, at this writing, is obtaining copyright permissions for digitization, server storage, and online use before the project begins, through our on-ground governmental contacts in Damascus and the Wizarat al-Adl (Ministry of Justice) Damascus, Syria, which holds the copyright. Discussions are in process with the appropriate government units.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

Using the Syrian Gazette digitization workflow as a model, our staff will then undertake
the digitization activities for the microform reformatting of the 30 years of laws and legislations
in the Palestine Gazette, from 1919 to 1948, in approximately 50,000 pages of various sizes.\textsuperscript{18}
Yale University Library holds this gazette in three microfilm groups: 1) under the title “Official
Gazette: Great Britain. Occupied Enemy Territory Administration (South)”, representing No. 1
(July 15, 1919)-no. 23 (June 16, 1920) in 1 microfilm reel (4 in., 35 mm); 2) “Official gazette of
the government of Palestine”, No. 24 (July 25, 1920)-no. 297 (Dec. 16, 1931) in 6 microfilm
reels ( 4 in., 35 mm); and 3) entitled “The Palestine gazette: official gazette of the government of
Palestine”, No. 298 (Jan. 1, 1932)-no. 1666 (Apr. 28, 1948) 33 microfilm reels (4 in., 35 mm).
All three were published in Jerusalem. The microfilm in each case was prepared by the Library
of Congress Photoduplication Service in 1979. The Gazette is printed in English and Arabic
over the course of its serial run. Yale’s Near East curator will work directly with Ministry of
Justice officials to secure permissions for the Palestine Gazette, as needed.

We recognize that both the proposed digitization and the research tasks that we describe
in this proposal will require significant time and effort. Since we further propose to establish
policies and to test methods for a repository architecture that is both interoperable and scalable,
we believe that in subsequent projects we will be able to complete the digitization of the entire
run of the Syrian Gazette and add the digital surrogates to the collection built during this project.

\textsuperscript{18}The pages of the material selected for digitization are A5 size in general. The international standard (ISO 216)
favors the most frequently used paper size, A4, which measures 210 × 297 mm. In the US (as well as Canada and
Mexico), the paper size Letter is the most common paper size. The Letter size is the familiar 8½ × 11 inches, or
exactly 215.9 × 279.4 mm. A5 paper measures 148 × 210 mm. We mention these standards to note that we must
address the standards when preparing images stored in the project’s repositories for printing.
3. Significance

National and International objectives

Since the Obama administration has taken office, significant diplomatic activity indicates a renewed desire to engage in dialogue and discuss regional challenges that have global consequence. There has been a marked expansion in contacts with Syria’s government. During March 2009, the State Department noted that these contacts involved expanding communications with important individuals in Syrian ministries, for example the Syrian Foreign Minister Walid Mu’allim, the presidential advisor on political and media affairs, Bouthaina Shaaban, and the Deputy Foreign Minister Dr. Faisal Mekdad. Clearly, key differences between our two governments remain, but the increase in meetings offers opportunities to discuss these concerns in open dialogue.

In March 2009, Secretary of State Hillary Rodham Clinton traveled to Egypt, Israel, and the Palestinian Territories, where she attended and participated in the donor’s conference for Gaza recovery hosted by Egypt. Special Envoy for Middle East Peace George Mitchell and other high-level representatives joined Mrs. Clinton on her first trip to the Middle East as Secretary of State. The continuing challenge of sustainable and durable peace in the region will require the good efforts of the Department of State and the Obama administration.

Access to materials covering pivotal points in the history of Syria and Palestine offer a valuable source of primary information with the potential to increase understanding about a time of nation building, applicable to today’s needs for urging Syria into a more productive international role while witnessing the democratic developments made by the Palestine National Authority.

19 For more details, please see: http://www.state.gov/p/nea/ci/sy/.
Theory, Knowledge, and Practices in related areas of study

The contents of foreign legislative collections are important to a wide range of researchers in the humanities and social sciences, as well as legal scholars and others. Areas of study such as ME politics, energy, migration, trade, investment, human rights and development increasingly require reference to foreign law. To meet these needs effectively, it is essential for researchers to have access to extensive runs of serial primary sources (especially official gazettes).

Syria has played and continues to play a pivotal role in modern ME history, but the absence of a large Syrian émigré community in the US – and other political factors – has kept it from receiving the focus that it deserves and that we need. The creation of Syria out of the body of the Ottoman Empire and its separation from and continuing relations with Lebanon, Iraq, and Jordan make it an enigmatic centerpiece to all ME study.

The Syrian Official Gazette (Al-Jaridah al-rasmiyah) is the official publication of the Syrian government. It was established in 1919, the same year that modern Syria emerged during the French Mandate. After four centuries of Ottoman domination, the Arabic speaking region of the Middle East was forged into a variety of modern states and entered a new era of independence and self-realization. The Syrian people enthusiastically supported the creation of a new Syrian state led by King Faisal, the son of the al-Sharif Husayn, who helped the Allied Forces in World War I to defeat the Ottoman Sultan. The youthful Syrian government was eager to make its country a modern state and wanted to see its people become a part of the new world century, which was being created from the ashes of the First World War.

The Syrian Gazette contains a wealth of information about modern Syria’s history and its social, economic, educational, legal, political, and cultural development, which cannot be found in any other archival sources locally or abroad. The issues of the Gazette record daily activities
of all the ministries, departments, and organs of the government, from the advertisements that were published for contracting a certain project, to political appointments and criminal convictions. Furthermore, this journal reported on new laws and official legislative matters. For example, in his book on Syrian rights, liberties and public duties, George Jabbour, former advisor to the late Hafez al-Assad, cited some 100 pages from the Syrian Gazette to describe the Peoples’ Council debate on the 1973 constitution. For historians and scholars, the Syrian Gazette is a cherished treasure trove of information for the study not only of modern Syria but also of other ME countries.

The official gazette we have selected is sparsely held, known, or accessible outside Syria itself – travel and work in Syria are difficult enough for scholars – and offers a unique primary source for the 20th century history of the country. It sheds much light on the history of early independent Syria in the 1920s and 30s, reflected in the work of scholars like Gelvin, Khoury, and others, as a contested but important contemporary source for the post-1967 emergence of Syria into a larger global political arena.

The Palestine Gazette (Official gazette of the government of Palestine) is the official gazette for the period covering the British Mandate as well as the Ottoman Empire Transferring Administration. A contemporary to the Syrian Gazette, it published its first issue in 1919 and ended in 1948 with the departure of the British. The terminating year is of utmost significance in Palestinian history, thus making the content contained in this gazette series of great scholarly value.

---


After the First World War, at the time when the Ottoman Empire was divided, the League of Nations issued the Mandate of Palestine. Initially, the exact boundaries of the Palestinian territory were undefined. Not until 1924, when the Allied Powers finally reached an agreement, did the Mandate specify its territorial limits. These extended from the Mediterranean to the western boundary of Mesopotamia (modern day Iraq), controlled by the British; to the north and northeast to the French-controlled Lebanon and Syria; and to the south and southeast the Kingdoms of Saudi Arabia and Egypt. In 1921, Churchill specified that the territory of Transjordan (today's Hashemite Kingdom of Jordan) was to be included in the Mandate but separated from an area for the establishment of a Jewish National Home. The Mandate continued until the catastrophic events of 1948 marked by the beginning of the Arab-Israeli War.

Dissemination of project materials

The late Edward Said, born in Jerusalem during the British Mandate, promoted a humanistic vision for a two-nation co-existence in the region once governed as the British Mandate. Said argued for the basic needs of the two ethnic groups – Palestinians and Israelis – such as food, water, and shelter. Access to the contents of this gazette would provide valuable primary source material to researchers – whether supporters or critics of writings such as those published by Said – about historical and legal decisions, events, and processes for a geographical region that today encompasses more than one sovereign nation. The contents, indexed and searchable, would represent a significant addition to resources for historians, humanists, and

---

22 The League of Nations Class A mandate under Article 22 took effect when the Ottoman Empire was divided in August 1920 by the Treaty of Sèvres following World War I. This mandate took place on April 25, 1920, at the San Remo conference in Italy.

Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

legal experts alike. On a broad level, project information will be disseminated by means of an assembly of experts. The project management team will consult on a regular basis with experts related to the selected materials. Among these are individuals and academic institutions dealing with scholarly issues for which access to digitized official gazettes can expand research capacities. (For more details, see Project Personnel section.)

Additionally, the digitized and openly accessible materials resulting from this project will complement historical and contemporary resources already available via the Internet. In preparation for this proposal, we conducted a survey using Google Books to determine the extent of availability of full text searchable content from monographs, government papers, and other publications. We tested for the presence of materials related to the League of Nations, Article 22. (The League of Nations Class A mandate under Article 22 took effect when the Ottoman Empire was divided in August 1920 by the Treaty of Sèvres following World War I. This mandate took place on April 25, 1920, at the San Remo conference in Italy.) The findings from Google Book searches show that many resources in Western languages from U.S. universities already exist in digital form on this subject thanks to the Google Digitization initiative. (See Appendix G for findings.) A similar search in WorldCat revealed several related monographs and journals, some in EBook format.24 We believe that a new digital collection of official gazettes from Syria and Palestine will supplement these materials well and will offer a more robust look at legislation at a time when nation building was most active. An added value will be linking the metadata tags from the gazettes to the keywords, place names, and subject heading that describe the Avalon Project collection, an archive combining both historical and contemporary documentation.

24 Searching the Library of Congress for similar keywords resulted in catalog entries, not digitized full text content, for constitutional and tax regulations.
On a project level, dissemination of information about the new digital collection of ME official gazettes will occur at the outset, during and at the completion of the project. Regular announcements will appear on the home pages of the project participants’ Web sites as well as be sent to the following:

Scholarly:
- MESA Bulletin: The Middle East Studies Association of North America’s journal of review
- MELANotes: The journal of the Middle East Librarians Association
- lis-middle-east: email list for MELCOM-International and MELCOM-UK

Technical:
- DigLib: A discussion list for digital libraries researchers and librarians
- Fedora Commons: The discussion list for Fedora developers, project managers, and institutional management
- University of Virginia Registry of Fedora repositories: a registry of Web-searchable digital collections using the Fedora framework

4. Project Design

The proposed project will begin on 1 October 2009 and finish on 30 September 2013. Effort on parallel time lines will address two major areas of focus: 1) collection building and digitization and, 2) research and technical sharing knowledge. (See Appendix H for a chart of tasks and milestones.)

Collection building and Digitization:
Yale will digitize and perform Optical Character Recognition (OCR) conversion on 70,000 pages in 59 oversize volumes plus six supplemental volumes of the Syrian Gazette, starting with the first volume published in 1919 and continuing through to 1948, a pivotal year in ME history. Next, we will coordinate the digitization activities for full text reformatting of the 30 years of laws and legislations in the Palestine Gazette, from 1910 to 1948, in approximately 50,000 pages in microfilm and print supplements of various sizes.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

**Research and Technical Knowledge Sharing:**

The technical lead at Yale will establish a wiki for communication among the team members and for soliciting community comments. A project web site will be 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 EAD mark-up of the gazettes, and begin the cataloging process for the various gazette formats. The project team will develop content models and metadata requirements and will oversee the automation of metadata creation and batch ingestion for the repository. Three lines of technical effort will continue from the second half of Year 1 through to the project end: 1) metadata generation activities including a Web-based tagging tool, 2) the archive presentation tasks including search and retrieval, page viewing, and citation generation, and 3) the integration of gazette content to other existing scholarly digital resources on the Middle East. An analysis of Avalon Project entries harvested via OAI or directly extracted from the collection will be the first point of integration of the newly formed collection with an existing resource.

**Milestones**

During **Year One**, the project team will select a vendor to provide services for scanning, processing, and OCR conversion. A team of skilled technicians, most with native Arabic skills, will manage the workflow tasks, including quality control checks. The members of the development team will configure a test copy of the Fedora framework to accept digital objects as these leave the final quality assurance point of the digitization workflow. The team will develop content models and metadata requirements and will oversee the automation of metadata creation and batch ingestion for the repository. A measure of success at the end of Year One will be the inclusion of at least 10% of the selected material in the newly configured digital repository using the content model, while approximately 30% of the digitization tasks will be completed. In
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

November of Year One, the project team will prepare press releases to announce the project and invite a review of the prototype under development by members of the academic community specializing in the ME.25

In the first half of Year Two, as more digital objects are added to the repository, the team will conduct a usability study of the prototype system to test the effectiveness of the search engine and the display features. Adjustments to the display features will occur based on study results. While digitization continues, the team will conduct tests on scripts developed to automate metadata creation. By the end of Year Two, the team will use the prototype metadata tool when depositing the remaining digital objects into the repository. ME specialists and archivists will review the finding aids created, based on accepted EAD standards. At least 50% of the selected material will be deposited into the newly configured digital repository using the content model, while approximately 70% of the digitization tasks will be completed. The technical team at Yale will begin conducting an analysis of the OCR-extracted text from the gazettes.

Year Three of the project will focus on digital production and testing of newly created tools. The team will submit the metadata creation tool for review by the experts assembled for consultation and advice. The team will augment all related documentation. The analysis of extracted text from the digitization process will continue by 1) creating a usable index from the analysis, and 2) creating the programming needed to insert cross-reference links. By the end of Year Three, the project team will begin investigations for further funding to continue the work of this project. The team will organize and hold a seminar on digitization methods on Yale’s campus for invited libraries.

25 This timing responds to the annual meeting of the Middle Eastern Studies Association which regularly meets in late November each year.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

In **Year Four**, OAI harvesting configuration for discovery of the new collection will be in place. The technical lead, working closely with the Yale Law School, will contrive the necessary structure for active links created from extracted text in the gazettes to connect to full-text entries in the Avalon Project. Tests to rate usability and simultaneous searching of materials held in the two archives will occur.

**Final Deliverables:**
Scanning, processing, and OCR conversion tasks will be completed by the end of the project term. We will have established and tested links from the initial set of existing scanned gazettes to the historical and contemporary materials held in the Avalon Project. 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, along with its findings regarding the new digital collection by sharing project documentation via the project web site for use by other academic libraries.

5. Project Personnel
The core team of project staff from Yale University Library brings a wealth of experience to the proposed project. (Please see Appendix I for curriculum vitae.)

**Principal Investigator:** **Ann Okerson** is Yale's Associate University Librarian with specific responsibility for Collections Development and Management and all international and area studies. She is particularly known, nationally and worldwide, for her work on the impact of electronic publishing, in the arena of electronic costs, and in copyright/licensing for the electronic environment (see http://www.library.yale.edu/~okerson/alo.html). In addition to the PI role, Okerson will provide support for intellectual property and economic issues.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

**Project Director:** Elizabeth A. S. Beaudin is the Manager of International Digital Special 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, during which time she has developed both academic and administrative systems in a university environment. Beaudin's work on the recent OACIS project included developing the software, interface, and technical communications for a new electronic union catalogue. She is project manager and technical lead for Project AMEEL and Iraq ReCollection, which is creating a scalable digital repository on the Middle East. Beaudin will coordinate all components of the project and will keep a day-to-day eye on progress toward meeting the specific project goals, as well as taking responsibility for leading the Yale team.

**Collection 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 will provide curatorial expertise and take the steps needed to secure copyright permissions for the selected materials. Mr. Samoeil has a broad and deep understanding of the field of ME Studies and the scholarly resources needed to support Yale's academic program. Before his arrival at Yale, Samoeil worked at the libraries of Harvard, the University of Pennsylvania, and King Fahd University in Saudi Arabia.

**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 an interface for use by team members and librarians.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

**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 University Library. Bogus will bring Preservation expertise to the project as the selected materials circulate to and from the outsourcing vendors.

**Digital Collections Specialist: Jennifer Weintraub** facilitates collection development aspects of digital conversion projects at Yale. Weintraub will represent the library’s Digital Production and Integration Program, a group composed of library specialists providing services for assessment and usability, digital consulting and referral, and digital production.

**Assembly of Experts**

In addition to the core team, the project management will consult on a regular basis with experts related to the selected materials. Among these are individuals and institutions bringing solid experience to the project.

**Scholars:**

- **George Jabbour** holds a PhD from American University and served in the Syrian parliament for many years. Dr. Jabbour has written books and many scholarly articles on Syrian history and politics.

- **Joshua M. Landis** is Director of the Center for Peace Studies at the University of Oklahoma and Assistant Professor of Middle Eastern Studies in the School of International and Area Studies.

- **Peter Sluglett**, the president of the Syrian Studies Association, is Professor of History at the University of Utah, who specializes in the history of the 19th and 20th century Arab Middle East, of the area which is now Iraq, Syria and Lebanon.

**Institutions:**

- **School of Oriental and African Studies (University of London, United Kingdom):** one of largest collections in world focusing on the ME.

- **Universitäts- und Landesbibliothek Sachsen-Anhalt (Halle-Wittenberg, Germany):** central institution for collection of ME digital materials in Germany.
Yale University Department of Near Eastern Languages and Literatures and Yale Council on ME Studies: department faculty and council members offer strong support that will take the form of scholarly expertise, guidance, and direction for the project as a whole.

Associations:
MELA (Middle East Librarians' Association) is an association of North American curators of ME collections offers its expertise as needed, access to the collections they represent, and guidance on the needs of users.

MELCOM (UK) was founded in the late 1960s as the Middle East Libraries Committee and maintains close ties with the British Society for Middle Eastern Studies (BRISMES) and MELCOM International, the European Association of Middle East Librarians.

MESA (Middle East Studies Association of North America), a learned society of international scholars, will serve as the voice of the project’s stakeholders.

6. Project Services
The project to unite in digital form the two official gazettes and the Avalon Project content will focus on utilizing a technical infrastructure and developing collaborative methods to support the digitized documents that have been created. By using experience gained from past projects at Yale, we will expand on an evolving technical architecture based on the Fedora repository framework to facilitate cross-collection indexing and searching. Further, we will focus efforts on solving the collection building challenges, such as best practices for microfilm reformattting, metadata creation, file transfer, and repository connectivity.

In addition, the Yale project team brings significant experience with Arabic text digitization. In 2005, the Yale project manager first completed digitization training at the Bibliotheca Alexandrina (BA) in Alexandria, Egypt. The staff of the Digital Lab at the BA is recognized for their expertise in Arabic text digitization, thanks to their participation in the
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

Million Book project. Their training and subsequent sharing of workflow practices enabled the Yale team to formulate a workflow adaptable for US research libraries. Since then, Yale OCR studies have shown that using robotic scanners increases accuracy rates, as does the use of font libraries when converting Arabic text with Sakhr’s Automatic Reader. (More software details are provided below.)

The methods we identify as part of this project will form a base for knowledge sharing in the technical community. Part of the challenge of this effort includes an analysis of international standards, in order to apply or adapt these when determining best practices for archiving and distributing newly digitized objects. We will pay special attention to documenting the microfilm reformatting methods used in the project.

There are inherent risks in any digitization project of this nature. For example, when performing an exhaustive review of the collection, it is possible to discover occasional missing pages or issues. Based on previous experience, we will use the network of academic libraries participating in the OACIS project, an electronic union list of journals from or about the ME, to locate holdings and to request missing issues via InterLibrary Loan. We will also consult the lists maintained by CRL and FLARE. In this way, we can create complete runs of the material to be digitized.

At the time of scanning the volumes, the binding, condition, or age of the paper, and the clarity of the printed text may have a bearing on the results, i.e. the quality of a scanned image may be lower than established standards dictate. The microfilm may have been scanned from imperfect originals. Because of poor source documents, the accuracy of full text created from

---

26 As of this writing, 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 permitting customized programming in three project teams to support all digitization efforts. Please see: http://dar.bibalex.org/#AboutDAR and http://www.ulib.org/index.html.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

OCR conversion may be lower than acceptable. Were this to occur in a significant amount of the materials selected for this project, alternate selection would need to be made from more modern or better-printed editions.

The following is a fuller description of methods and techniques for material processing, organization, and access to the newly digitized material, as well as maintenance and longevity of the project and its contents.

Material Processing

As mentioned above, in Yale University's Project AMEEL digitized journals published in the ME are the target content for a new searchable Fedora repository. The experience gained in designing the technical architecture and configuring a digitization workflow will be applied to the G2G project, specifically because of the immense challenges related to the digitization of Arabic text. Our Gazettes project thus builds on and complements this earlier investment in Arabic digitization. The preparatory and processing steps described below reflect lessons learned when designing a workflow at Yale for the digitization of Arabic text.

Shelf Checking and Workflow

Continuing with the initial shelf checks done in preparation for this proposal, the digitization team will note physical characteristics of each volume. (Please see Appendix J for a sample shelf-check form.) They will also note the language of publication, the contents of the volume including starting and ending page numbers, presence of front and back matter, table of contents, and other elements. The team will update a workflow control database using the shelf check information. The directory structure and image naming requirements for workflow control are based on the name for each gazette, volume number, and language of publication. This last element is an essential part of the file name since it indicates during the workflow where OCR
software for Arabic or for Western scripts is used. Since we plan to send the majority of the selected material to outsourcing vendors for scanning, an outsourcing control report will be prepared for each shipment of volumes to be sent to the vendor based on data collected in the workflow database. (Please see Appendix K for a sample of Yale’s outsourcing controls.) At times, due to binding or other physical conditions, selected volumes will be scanned and processed in house based on the workflow developed at Yale during the AMEEL project.

Standards

Our standards indicate that the printed material for digitization should remain bound during the scanning.27 By using outsourcing agencies providing scanning services, we will produce digital equivalents of each page, scanned at 300 dpi, bitonal for text, and greyscale for images.28 These settings for the scans have proven most successful in terms of the OCR requirements for the language of the selected material. Specific requirements for creating a faithful reproduction of the source material will be adjusted as needed, depending on the age and condition of the material. The output format, conforming to recognized best practices, will be TIFF (Tagged Image File Format).29 The bitonal TIFFs will be compressed using CCITT Group IV and the greyscale TIFFs can be compressed using LZW.30 Once scanning is complete, the

27 The Yale team sent out ‘Requests For Proposals’ (RFPs) as preparation for this proposal to three vendors. Via word of mouth, five vendors have sent in RFPs for consideration. With this information, the team can move rapidly to make a selection. The costs estimated in the proposed budget are based on the results of the call for RFPs.
29 TIFF is widely acknowledged as the best format for preservation and technical longevity.
30 LZW is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch, and published by Welch in 1984. CCITT Group IV 2D compression supports Black & White images. CCITT comes from the French name “Comité Consultatif International Téléphonique et Télégraphique”), a telecommunications standard created in 1956.
processing phase involves image clean up such as the standardization of margins, de-skewing, removal of noise on the page, and resizing the image canvas to a uniform size.

Regarding standards for digitization and OCR of microfilm, we will consult first with the Northeast Document Conversion Center. (http://www.nedcc.org/home.php) As in previous projects, we will adhere to existing standards and carefully document results during the project so that we may publish our findings periodically and at the end of the project.

**Optical Character Recognition (OCR)**

Three OCR software products will be used: 1) Automatic Reader – OCR Platinum 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.

31 Other sources for guidelines include the Library of Congress (http://www.loc.gov/ndnp/digitizingmfocrlinks.html) and IFLANet (http://www.ifla.org/VII/s39/broch/microfilming.htm).
Quality Assurance

Quality control steps will occur at several intervals. First, we will prepare an outsourcing control report of the materials being sent for scanning. When these materials are returned from scanning, we will check for the completeness of the return shipment and the condition of the volumes. In addition, we will verify the output files received against the outsourcing control report. If more than 2% are not up to standard, we will ask the vendor to redo the entire batch. Output from OCR tasks will be checked for completeness and accuracy. We will ask the outsourcing vendor to perform: 1) a visual Quality Control check to ensure that all .txt files are properly encoded as Unicode UTF-8 and display Arabic characters, and 2) a random sampling of 10% of the total batch of files, which will be reviewed for OCR accuracy. Arabic OCR output accuracy levels are expected to exceed 90% on modern texts and our goal is to reach 96.5% – a very high standard when dealing with Arabic text.\footnote{We expect and have witnessed accuracy rates much higher than 90%, usually reaching 96.5%. When selecting an outsourcing firm for OCR tasks, we will verify that samples provided prior to engaging the firm and delivered work match our expected standards of accuracy. We state that levels “are expected to exceed 90%” to acknowledge that some materials in the selected collections may not yield the higher rates due to physical conditions such as age and quality of ink used at the time of publication. Sakhr’s Automatic Reader can produce OCR accuracy in excess of 98%. In OCR tests conducted at Yale, robotic scanning and automated image enhancement increased accuracy ranges to 98.25%. In an earlier study at Yale using the BA’s method of calculating OCR accuracy (software-provided statistics plus manual error counting), manual scanning and the use of font libraries produced accuracy rates of 94.4 to 96.0%. A significant difference in these two studies results from the scanner in use: a Minolta PS 7000 Black & White book scanner does not perform as well as the Indus 5002 Color book scanner now in use at Yale.} While it is commonplace for English-language OCR to approach 100% accuracy, the Arabic OCR software packages achieve initial results of 80 – 85% and require a great deal of "learning" and adjustment to exceed 90% or better – and to do this in an "automated" way; our project will contribute to this type of research. The materials selected for digitization provide an added challenge to optimizing OCR accuracy in that there is a mix of languages in both gazettes, i.e. Arabic and French in the Syrian gazette, and Arabic and English in the Palestine Gazette.
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

We will perform a secondary quality control check 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.

**Organization and Access**

The architecture for the gazettes collection will use the Fedora framework repository software residing on a Linux server (2.6.9). The archival TIFFs are stored on a separate server. The Fedora framework provides for OAI compatible harvesting for resource discovery, thus permitting other repositories to discover G2G metadata. Within the Fedora framework, a basic search function is provided. In addition, G2G will use the Fedora General Search service and customized searching adapted from the open source product called AraMorph.

Among other technical features, the gazettes collections will make use of and adapt the content model for adding new digital objects to the repository, previously tested and used at Yale. This content model controls the hierarchical relationships between such object elements as title, volume, and page. These content model rules help to define how digital objects will be retrieved and displayed as the result of a patron search. The use of persistent IDs for all digital objects will be thoroughly reviewed and verified for consistent access to the digital material. In this way, the project team will examine how to handle different formats in the interests of flexibility and scalability for accepting partners in the future.

In terms of metadata, we will begin with metadata records created initially by selecting the Online Public Access Catalog (OPAC) record for each title. This project will provide a

---

33 For more technical information, please see: [http://www.fedora.info](http://www.fedora.info) and [http://www.fedora-commons.org/](http://www.fedora-commons.org/).
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

MARC XML file at the title level for each gazette. This file will be available at viewing time. At each level of the content model, a customized Dublin Core (DC) file is added at the time of ingestion. (Please see Appendix L for sample Dublin Core metadata). The DC schema includes qualified DC elements for richer searching and for customizable displays of access rights information. The MARC 21 format, as documented by the Library of Congress, and DC standards will provide common standards that can be used to unite future collections on an international level. DC citations in the digital records will also follow the OAI protocol, thus enabling OAI harvesters on the Internet to locate subjects covered in the gazette content. In addition, the emerging standards from the National Information Standards Organization (NISO) will be used when creating the structure of the database and its digital objects.34

Team members, with the necessary linguistic and technical talents, will conduct an analysis of metadata generation needs and work on the development of an automated metadata creation system. This process will include analysis of AraMorph to draw up specifications for the generation of bilingual metadata and finding aids.35 Additionally, we will review work done by the Society of American Archivists, especially as this relates to existing tools.36

Patrons will gain access to the G2G collection via a specially branded portal. In addition, patrons can discover the contents of the collection from at least two different vantage points, that is, the repository discovery tools specific to Yale Library and the Law School. Indexing and searching provided by the Fedora framework, full text content, and bilingual finding aids will give patrons a listing of results based on their search criteria. Access to the full text of search

---

35 For best practices on finding aids using Encoded Archival Description (ead), we have consulted the Library of Congress site at: http://www.loc.gov/ead/ .
36 See http://www.archivists.org/saagroups/ead/ .
results assumes that the digital objects adhere to copyright regulations in the US and countries of origin. If the administrative metadata of a particular object are coded with less than full rights, the patron will still be able to conduct a search and see a listing of search results.

Throughout the project, we 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 Linux servers used at Yale include Enterprise software managed by Red Hat and include systems management subscriptions to keep the server software protected and current. Systems support staff at Yale University Library regularly monitor server traffic, guard against attacks, and apply widely accepted security practices to development and maintenance servers.

7. Resources

The Yale project team has gained considerable expertise in Arabic text digitization. Starting in 2005 with the assistance from the digital staff at the Bibliotheca Alexandrina, the Project Director learned the essentials of Arabic text digitization. In 2006, the Project Director and Yale’s Near East Curator participated in collaboration with Stanford University Library and the Bibliotheca Alexandrina in a Hewlett-funded planning grant called DigiArab with the aim of digitizing 50,000 books printed in Arabic. In 2007, the Yale team sponsored a 5-day digitization workshop in Alexandria, Egypt attended by a group of 14 librarians, publishers, and vendors. The sessions included lectures and hands-on practice on scanning, processing, and Arabic OCR conversion based on the workflow at the BA. In June 2008, the Project Director delivered a talk on AMEEL and Arabic text digitization at the annual Middle East Librarians Committee (MELCOM) conference at St. Antony’s College, Oxford. At the November 2008 annual Middle

37 For more details on the Enterprise management entitlements, please see http://www.redhat.com/.
38 For more on IT security policies at Yale, please see: http://www.yale.edu/its/policy/
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

Eastern Studies Association (MESA) conference held in Washington, the AMEEL team held a workshop in conjunction with the American Association of Arabic Teachers (AATA) on managing Arabic text digitization projects geared toward the needs of U.S. academic libraries. There were 28 participants including faculty, librarians, and technicians. The Yale team also sponsored a special panel session to discuss the AMEEL project from different professional perspectives. We aim to maximize the lessons learned in these opportunities by working to integrate materials held in the Avalon Project at Yale Law School and by learning the intricacies of reformatting microfilm and its relation to OCR conversion of text.

Yale University is bringing some of its best talent to bear on this project and is providing cost-sharing at and above (on a year-by-year basis) the 33% minimum required for funding applications through TICFIA. The Budget Narrative Section outlines detailed costs and cost-shares. It should be emphasized that the opportunity costs for Yale are substantial as well, in that our commitment to this project concentrates resources that would otherwise be distributed among other continuing commitments.

8. Project Management

Commitment of the project team

The project team has collaborated on the selection of the materials and has discussed the preliminary technical challenges. There is a keen desire to make this valuable collection accessible to researchers seeking primary source materials on two areas of recognized national and international interest. The team has committed to the time and effort required to do so according to accepted standards and to a high degree of usability.

39 http://www.library.yale.edu/ameel/MESAworkshop/index.htm
Proposal to TICFIA: Yale University’s Gateway to Gazettes (G2G)

**Sustainability**

Yale University Library is committed to hosting and sustaining the G2G project as another in a series of signature projects of the University's commitment to learning and understanding in Arabic and ME studies and as a contribution to the common enterprise of scholars and other users around the world. The project team will have built on the experience gained in three previous projects: 1) OACIS, an electronic union catalog of serials from or about the Middle East; 2) Iraq ReCollection, a digitization project compiling full text content for nine scholarly journals from Iraq; and 3) AMEEL – Arabic and Middle Eastern Electronic Library, a scalable full text repository and portal providing scholars with centralized access to academic materials on the ME.

It is essential for any granted project to create a product that is not only scalable in design but that also can be supported and used after initial development is completed. For this reason, the project team will also focus on sustainability issues of two sorts – financial and technical. In the first case, future funding from granting agencies, as well as from commercial and private sources, will be sought in conjunction with the sustainability planning of Project AMEEL. In the second case, the project team will use follow accepted digital preservation standards to ensure the technical continuity of the materials digitized, that is, in terms of the life of the storage media and the accessibility to the contents of the digital collection. Yale University commits to ongoing maintenance of the server and its digital collections.

**9. Evaluation**

The purpose for conducting an evaluation of this project is not only to meet the requirements of the funding agency for accountability of the expenditures of the grant, but also to provide management information regarding the sharing of digital resources and the creation of new metadata generation tools. Because we propose a significant research and development
component, the project evaluation will aim to gauge the following: 1) the efficacy of the transfer methods used during the project (since data transfer methods will be used throughout the project), 2) the success of the new metadata tool, and 3) the compilation of test results from searches conducted by a control group of students on selected university campuses.

**Data transfer methods:**

Careful tracking of the data transfer methods will be part of the grant since we plan to make use of outsourcers, when possible, to keep scanning costs low. At both the midpoint and the end of the grant, we will compare the number of files transferred and the time needed to ingest transferred files into a repository. Success will be determined by a 50% improvement rate from transfer date to ingestion date. To measure improvement, we will first exchange workflows by the beginning of January 2010. The team will analyze the workflows with the selected scanning vendor to suggest where time might be optimized. At the end of the grant, we will evaluate the statistics to a) report whether or not a workflow could be optimized and b) disseminate to others our most effective work plans.

**Metadata and finding aids tool:**

Yale has significant prior experience with metadata use in digital libraries holding books and journals. We wish to determine how these models can be applied to the gazette format. We plan to share this evaluative information, especially since a content model for gazettes does not yet exist. For example, extensive work with Fedora and metadata schema done at other institutions does not address gazettes as a specific content format.40

The metadata generation tool will be considered a success when staff can enter pertinent data and review this data for completeness. We will determine completeness by documenting basic metadata elements at the beginning of the project. We will track the time needed to prepare basic metadata elements as new digital objects are added. A group of selected

---

40 See: [http://wiki.dlib.indiana.edu/confluence/display/INF/Fedora+Content+Models](http://wiki.dlib.indiana.edu/confluence/display/INF/Fedora+Content+Models)
stakeholders and experts will review these elements in Years Two and Three of the grant and report to the project team. The team will report on the findings in relationship to the time needed to generate the metadata.

**Cross-repository searching:**

Yale University will develop and conduct a usability study using small groups of students along with experts identified for this project. The study will focus on specific types of searches that retrieve data from each repository separately and the newly formed joint collection. We will judge success when testers can retrieve 80% or more of the search materials. In addition, we seek to measure how the newly formed collection is used in order to illustrate project value to TICFIA and others. For this reason, we propose to record the number of hits per month to the front page, i.e. the front door for searching content, and to other points of entry such as web pages maintained by Yale Library and the Yale Law School. Further, we will track the origin of page hits by aggregating the network and subnet components of patrons’ IP addresses. From the data gathered, we will deploy monthly reports to the project management team, quarterly reports by means of the project’s web site, and yearly findings in TICFIA reports.

Results describing these evaluation points will be shared in writing in the end of grant report and as part of a white paper to be disseminated as described in the grant proposal. We will include workflow diagrams in our documentation and a critique of the work plans conducted by the project team.

---

41 An IP Address is defined as an Internet Protocol Address, or the number assigned to a device in a network utilizing the Internet, e.g. 130.132.111.xxx, the network and subnet components – 130.132. – can locate the origin of the inquiry. To protect individuals’ privacy, we will not track the final two components.