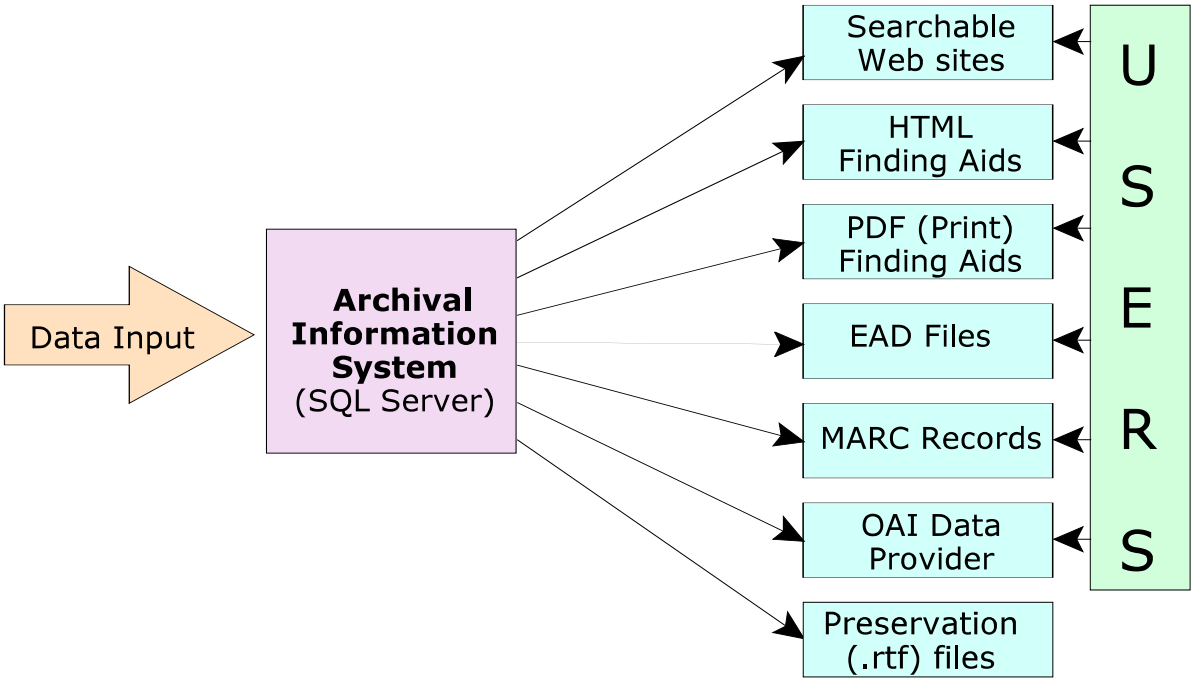
A Unified, Standards-Compliant System
for Describing Archives and Manuscript Collections

Christopher J. Prom, Assistant University Archivist
Scott W. Schwartz, Archivist for Music and Fine Arts

Amount Requested: $9,260
Project Period: January 15 to August 30, 2005



ABSTRACT: We request ICR seed funding to hire a computer programmer to work full time for three months during the summer 2005. The programmer, Chris Rishel, will assist us in prototyping an innovative approach to providing record series/collection, sub-series, and file-level control over archival records and manuscripts. Work completed during the spring and summer of 2005 will support an IMLS National Leadership Grant application which we plan to submit in February 2006. Funding will prepare the system for production-level implementation under the IMLS proposal, which we anticipate will fall in the range of $100,000-150,000.

The seed money will be used for database design and output modeling. The prototype system will provide common a platform for data management, comply with all relevant standards for describing archival materials, and provide robust data interoperability. Archival descriptive records will be exported as HTML, Encoded Archival Description (EAD) files, MARC catalog records, PDF files, OAI records, and Rich Text Format files. The system will dynamically update collection-level MARC records in Voyager and EAD files in RLG's Archival Resources. Tools developed under this seed money proposal and the IMLS grant will be made available as open source software.

PROJECT BACKGROUND, SIGNIFICANCE, AND RATIONALE

The University of Illinois Archives, Sousa Archives and Center for American Music, and American Library Association Archives currently manage descriptive information regarding archives and manuscript collections in three separate databases. These databases provide dynamic access for primary source research by campus students, faculty, and administrators as well as external scholars and members of the public. Web interfaces are available at:

http://web.library.uiuc.edu/ahx/uaccard/default.asp
http://web.library.uiuc.edu/ahx/ala/ccard/default.asp
http://www.library.uiuc.edu/sousa/index.php?p=finding#

The three databases were constructed using funding from our units. In the case of the University Archives and ALA databases, they were converted by Chris Prom and undergraduate assistants from off-line systems originally developed by Maynard Brichford and William Maher in the 1970's and 80s. The Sousa Archives database was prototyped by Chris Rishel, an undergraduate computer science major working under Scott Schwartz's direction. While the three systems currently provide on-line access, they need to be upgraded in several key respects.

The money requested in this proposal would fund technical work to merge the three databases. It will help us prepare the system for production-level implementation, allowing us to convert data from the proprietary WordPerfect format and develop open source tools. This work cannot be completed using existing resources since it involves significant technical expertise and programming beyond the capability of current staff. The project is also too large to be completed by systems office personnel, although the systems office, the Digital Services and Development Unit and the Digital Content Creation Team will be consulted actively during the project. These units/groups have advised us as we developed this proposal.

The proposed system will hold many forward-looking design features, making it an appropriate model for IMLS National Leadership Grant funding under the Building Digital Resources rubric.

1. Complies with Descriptive Standards

At this time, the systems we intend to merge use proprietary file formats and lack key data elements. As such, they do not comply with best practices for describing archival materials. This is common to the descriptive systems at many other academic archives, and the products of the IMLS grant we intend to submit after completion of this pilot project will provide models and tools that can be adopted elsewhere.

The current University Archives and American Library Association Archives databases were created by merging data converted from mainframe and local systems originally developed in the

1970s and 80s.[1] Inherent limitations of database size and structure mandated a simplified design. These features were replicated in our current system. For example, fields provided for the University Archives and ALA Archives include only Record Group, Subgroup, Record Series Title, Inclusive Dates, Arrangement, Volume, Description (Scope and Contents), Subjects, and Date Received. While these are sufficient to generate a searchable website, the current systems cannot disperse descriptive information to other user access systems. Important data fields, such as creator, are not included. Furthermore, the UA and ALA databases do not support the ability to record descriptive information for box and folder lists, except as a link to a PDF file created from a data in a proprietary file format (WordPerfect).

The Sousa prototype database similarly includes support for some but not all of the elements called for by the descriptive standards described. For example, it does not include support for subject indexing. However, unlike the University Archives database, sub-series and folder titles can be tracked. Furthermore, the data model undergirding the database is not robust enough to allow for dynamic updates of box/folder information as new materials are added.

The new system would merge the best elements of the three databases while adding functionality not currently provided. Relevant standards that the new system will support include:

- *General International Standard Archival Description*.[2] The system will include data elements recommended by ISAD (G). This standard provides general guidance for preparing descriptive records for archival materials. Archival description provides information about the context in which records or manuscripts were created (i.e. their provenance), as a well as a description of the content of the materials themselves. ISAD (G) discusses the need for "multilevel description" and provides 26 elements for describing archival materials. Sample elements to be added to our system include creator, biographical history, appraisal/destruction information, and related materials.
- MARC21. The system will dynamically update MARC records in Voyager. The MARC21 standard includes support for the MARC format for Archives and Manuscripts Control.[3] MARC-AMC was developed in the early 1980s by the Society of American Archivists to provide a structure within which archival repositories can produce bibliographic descriptions of their holdings. These records can then be included in standard library catalogs. MARC AMC does not allow sufficient flexibility to relate collections or parts of a collection to each other, so while it is one exchange format, we cannot rely on it to bear the weight of the overall descriptive system.

---

[1]Prior to this, descriptive information was held in paper or word processed documents.

[2]http://www.ica.org/biblio/cds/isad_g_2e.pdf

[3]Richard P. Smiraglia, *Describing archival materials: the use of the MARC AMC format*. Haworth Press: 1990.

- Encoded Archival Description (EAD).[4] The system will automatically produce and update EAD files in response to changes in the database records. Encoded Archival Description is a data structure standard for the recording and exchange of descriptive information about archival materials.
- *Describing Archives: A Content Standard.* DACS is a content standard for the construction of archival descriptions. As such it is the archival correlate to and extension of AACR2. It provides rules by which the content of data elements (for example, subject entries or creator names) should be constructed.[5] Compliance with DACS cannot be mandated by the computerized aspects of the system. Nevertheless, the database will provide a common platform under which DACS-compliant EAD files and MARC records can be produced. We will consult with cataloging staff to develop simple implementation instructions to be used by archival staff and graduate assistants creating descriptive records. Such rules will ensure that records created in the future will comply with the DACS recommendations to the greatest extent possible given staffing resources.

2. Uses Innovative Data Model

The system we propose will be an integrated system allowing searching both within single series/collections as well as across entire series/collections of material. It will use SQL server as a data storage medium, but be able to generate fully compliant MARC records, EAD files, and other formats for data exchange. Prior literature demonstrates that it is possible to map archival databases to MARC format.[6] After consulting with John Weible, Michael Norman, and Chait Naun Chew, we have identified a method by which Voyager records can be kept current with our database. EAD files can be produced using a similar method; such records have already been created in an experimental fashion for the Sousa Archives database. (EAD would be used in this system as an exchange format only, similar to the MARC file.) Furthermore, the system will include support for name authorities and controlled subject headings constructed according to the rules mandated in DACS.

The systems we propose to replace are currently stored as separate but parallel files on the library's production database server, Microsoft SQL Server. Our units maintain three sets of data, three input mechanisms, and three sets of code for producing webpages. The systems are inefficient because the database structures track the same types of information. Database

---

[4] http://www.loc.gov/ead/

[5] *Describing Archives: A Content Standard* (Chicago: Society of American Archivists, 2004). http://www.archivists.org/catalog/pubDetail.asp?objectID=1279

[6] Ronald J. Zboray, "dBase III Plus and the MARC AMC format: problems and possibilities," *The American Archivist* 50 (Spring 1987): 210-25. It should be noted that without the intervention of a cataloger or very tight compliance with content standards, such records may not be standardized enough to exchange with OCLC or other national systems.

maintenance issues are extremely time consuming. Merging the files to a common and well-documented database is the obvious solution to these problems, although the system will need to be very well designed to avoid bugs and maintenance issues.

In the past, we experimented with manually creating EAD finding aids and MARC records. Many academic archives currently use a similar model or are simply producing print finding aids. But our experiments have also convinced us that creating files which are stored in the native MARC and EAD formats is not supportable. It is time consuming and costly. It requires a great deal of training. File maintenance is extremely difficult if not impossible. It is a waste of precious staff time and resources creating records that are difficult to manage, modify, and re-purpose.

The system we envision would have a common data input mechanism (a web browser) and storage medium (SQL server), but allow for the export of data in many formats: MARC, EAD, OAI, dynamic (searchable) HTML, PDF files, and potentially other exchange formats as they are developed. Data that is input once would be output in many different ways. Similarly, corrections to data files will take place at one access point, but changes will propagate to records exported to multiple locations and accessed by multiple service providers and end users. For example, EAD files donated to RLG and MARC records in Voyager will automatically update each time a change is made to the underlying database record. Prototype work done by Scott Schwartz and Chris Rishel has successfully created EAD and PDF files from the Sousa database. Work on this phase of the project would apply the prototype in a broader environment and test its scalability.

3. Supports Folder-Level Description

As noted above, the University Archives and ALA Archives databases track archival descriptive data only at a record series/collection level, not directly at a sub-series or file level. As a result, the Archives commonly prepares print finding aids using word processing software. While this provides access to the folder lists via linked PDF files, it does not present descriptive information in preferred formats, such as EAD or as dynamic, searchable websites.

The proposed system will record descriptive information regarding the sub-series, and file folders within a collection, building upon the prototype work completed for the Sousa Archives. The system will include "triggers" that will allow for box lists to be updated easily and quickly as new materials are added to record series/collections. Access to folder-level information in multiple formats will enhance end-user access by allowing us disperse information to multiple communities.

In the prototype phase of this project (i.e. the work completed with ICR money), the system will support only three levels of description: record series/collection, subseries, and folder level. Supporting these minimal levels of description will provide a proof of concept for expansion to additional levels in the complete IMLS grant proposal, including subseries and item level. The

final IMLS proposal will also include a request for funding to convert selected WordPerfect finding aids pages into the system.

4. Facilitates Multiple Access Points

Adding missing data fields and allowing the database to track folder titles will enhance access to archival record series and manuscript collections. In addition, the money requested in this proposal would allow us to develop replacements for the current web sites. The current interfaces, while usable, have several significant usability problems which I (Chris) identified in a recent research study.[7] These problems would be eliminated when the user interface is redesigned. The Sousa Archives website will also incorporate these system-wide improvements.

Currently, our users find out about our collections only through our website. End-user access to the collections will be expanded dramatically because bibliographic records will be available in Voyager and EAD files will be made available in national databases such as RLG's Archival Resources. Not only will providing a record in the catalog and the RLG database better comply with best practices, it will open up access to people using standard library tools to locate archival materials. Finally, the system would allow us the option of exporting MARC records to WorldCat. While the MARC records would need to be carefully vetted by Archives staff prior to upload, John Weible and Chait Naun Chew indicated that it should be possible to design a system that will allow us to transfer the records on demand.

In addition, the IMLS project will include the development of dynamic websites which allow searching within a single collection, across all collections, or within a user-selected subset of collections. One of the biggest disadvantages of using EAD as a storage format is that tools for developing cross finding aid searches (such as DLXS) are difficult to implement and generally present archival information in a manner that is difficult to customize and search. Using EAD alone, there is no simple way to preserve relationships between collections of materials. On the other hand, the use of a standard scripting language such as PHP to query a SQL database will provide complete flexibility in the output mechanism and allow for a cleaner interface.

5. Develops Open Source Software

The software to be developed under the final IMLS proposal will include tools written using PHP. All tools will be provided through Source Forge for implementation by other institutions. Although the database will be stored in Microsoft SQL Server, complete documentation for the database structure, data relationships, and other technical matters will be provided on an open source basis under the terms of the General Public License or a similar license.

---

[7] Christopher J. Prom, "User Interactions with Electronic Finding Aids in a Controlled Setting," Forthcoming, *American Archivist*. Draft available at http://web.library.uiuc.edu/ahx/workpap/interactions.pdf

6. Lays Foundation for Digital Archives

The design chosen for this project will provide the technical underpinnings for digital archives projects which we hope to undertake in the near future. For example, an Archives Photo Database is under development using separate funding. A digitization project for the Board of Trustees Proceedings is also underway. Digitization of Sousa scores is also planned. The projects share metadata elements with the system proposed here. The robust system design will allow direct access to digitized or born digital materials on our website or in other resources, such as archival content placed in UIUC's planned digital repository. By the same token, the existence of this system will allow the metadata for digital objects to be designed in such a way that they include links back to record series/collection descriptions.[8]

SCOPE OF WORK/TIMELINE

The work to be undertaken during the seed grant period involves essentially three steps. This work would create a prototype of the initial database, provide models for the data output, and build support for the conversion of WordPerfect finding aids during the full grant period. As these three steps are implemented, work will be reviewed by the University Archivist, systems office staff, and catalogers who are listed below as project consultants.

Database design and data conversion: Prior to the project start date, Chris Prom and Scott Schwartz will analyze the current database table designs, data structures and field definitions, cross referencing current tables and correlating fields among databases. Tables and fields will be renamed to adopt naming conventions and ease future data maintenance. After the project start date, data fields will be added to the University Archives database to allow for smooth import and merging of the data, so that data from the Sousa database is not lost during import. In addition, support for controlled vocabularies and the Library of Congress name authorities will be added. Data from the University Archives, ALA and Sousa databases will then be imported into the new database. A series of checks will be completed to verify the accuracy of data conversion. A complete set of documentation will be created including a description of each data table, its relationship to other tables, and a description of each filed in the tables. This documentation will be maintained on the project website. **To be completed May 15-June 15**.

***Input Mechanism Design*:** In spring 2005, Chris Prom and Scott Schwartz will develop a list of functions that the data input mechanism should support. After consulting with Chris Rishel, systems office staff, the Digital Services Development Unit and the Digital Content Creation Team, we will supply specific directions for developing the input tools. We will oversee the process of programming and test the input tools as they are created. User documentation will be written. **To be completed June 15-July15th**.

---

[8] These records may be provided as METS and MODS objects or potentially other digital resources.

Outputs Programming: Anticipated outputs from the system include a searchable website (in PHP), EAD files, MARC Records, OAI data provider records, PDF files, and rich text format files.[9] In the spring of 2005, Chris Prom will map the database fields to the analogs in each of these formats, using prototypes previously completed for the Sousa Archives database. (During the ICR grant period we anticipate that only enough time will be available to prototype new web sites and develop the scripts to export text files. For reference purposes, we list below all of the work to eventually be completed under the IMLS grant.)

> For the ***dynamic HTML websites*** and ***rich text format files***, the existing University Archives, ALA Archives, and SACAM Archives web sites will redesigned and RTF files will be output using PHP scripts using criteria to be supplied by Chris Prom and Scott Schwartz. If necessary, Chris Prom will complete some of the programming for these within the ICR seed money period. For the ***MARC records***, the mapping will be vetted by Michael Norman and Chait Naun Chew. The programmer will work with Systems Office staff and Naun Chew Chait to develop the connection between the database and Voyager. A script will be developed to allow us to export selected records to OCLC for inclusion in World Cat. For ***EAD files***, we will use experience gained in implementing the Sousa prototype project. A template mapping database columns to output fields will be supplied to the programmer, and he will write PHP scripts that generate EAD files on demand, placing the proper information from the database fields into the EAD files. For the ***OAI data provider***, Chris Prom will supply an updated mapping to Tom Habing, who maintains the OAI data provider for the archives databases. Tom will update the current provider site to include the new information. For ***PDF files***, PHP scripts will be written by the programmer, working under our supervision and in consultation with systems office and DSD staff. Working under Scott Schwartz's direction, the programmer is currently protoyping scripts which produce PDF files on demand for the Sousa database.

Work on dynamic HTML and text files will be completed between July 15-August 15, remainder of work to be completed under IMLS grant.

PERSONNEL/BUDGET

Based on our past experience working with similar projects, we are confident this work proposed in this ICR seed money proposal can be completed by one programmer working full time during the summer of 2005. Chris Rishel, an undergraduate employee who has worked on the Sousa Archives website under Scott's direction, is sufficiently skilled to undertake the work.

Rishel bring several key abilities and experiences to bear. He is a skilled programmer well versed in the programming languages necessary to undertake the project. He has direct experience working with the Sousa database and in that context has prototyped some of the

---

[9] If feasible, we may experiment with creating formatted text output in Open Office file format. Open office uses an XML file format. http://www.openoffice.org/

innovative functions which our system will support. He is familiar with archival materials, receptive to learning about the archival functions the database supports, and understands the EAD data standard. He is a James scholar enrolled in the computer science program and has direct experience with designing web interfaces for both the Sousa Archives and the Illinois Department of Aging, where he designed a database to be managed and accessed by managers and line staff. He has provided a firm commitment to the project at a wage of $18.50 per hour. We believe this is fair given the technical complexity of the work as well as the strong past performance we have seen. A copy of his resume is attached for reference.

In addition, the following Library staff have agreed to serve as project consultants on an as-needed basis as we prepare for an IMLS grant:

- William Maher– project review, archival functions, staff workflow
- Michael Norman–Voyager, MARC mapping
- Chait Naun Chew–Voyager, MARC
- Tom Habing–date interoperability, OAI
- Robert Manaster–programming support, SQL Server
- Nuala Koetter–digital library design, metadata compatibility
- Kathleen Kern–public services issues, user interface
- Beth Sandore–IMLS grant writing assistance

Wage for programmer: $18.50/hour.	13 weeks @ 40 hours per week:	$9,620
	Total	**$9,260**

FUTURE PLANS

If this proposal for ICR seed money is accepted, we will undertake the following steps between now and the anticipated IMLS grant submission deadline of February 1, 2006:

- Completion of work outlined above
- Literature review in areas of interoperability and archival descriptive systems
- Environmental scan to identify similar projects undergoing, such as the Mellon funded "archivist toolkit" project.[10]
- Consult University Archivist, archives staff, and users for design requirements
- Draft white paper discussing functional objectives of system and laying out basic design elements.
- Undertake planning for the ICR work to be completed during summer of 2004, as outlined above.
- Contact Martha Crawley at IMLS to discuss potential grant and project design.
- Contact additional co-PI's to serve on IMLS grant (Possibly Nuala Koetter, Tim Cole, Carole Palmer.)

[10] http://euterpe.bobst.nyu.edu/toolkit/

- Arrange project consultant for IMLS grant (possibly Michael Fox, Minnesota Historical Society)
- Solicit letters of support and commitment for IMLS grant
- Develop full grant application, including narrative, workplan, budget, and cost sharing models, reference work to be completed in the seed money grant as basis for additional work to be completed in main grant proposal.

Expected outcomes for IMLS National Leadership Grant:

- Production-level system managing and exporting archival descriptive information at collection, series, subseries, file and item level
- Export of archival information in multiple formats: dynamic HTML, PDF, EAD, MARC, OAI.
- Upgrade of existing records in University Archives database to compliance with ISAD (G) and DACS standards.
- Linking to digital archival objects.
- Incorporation of archival information in Voyager, WorldCat, RLG's Archival Resources and other repositories.
- Easy to implement system for generating standards-compliant descriptive records from academic archives.
- Open-source software for application in other institutions.
- Technical experience and data model for potential regional (Illinois or Midwest) archival information service.

SUMMARY

Developing this project first as an ICR seed grant then as a full IMLS proposal is a logical step toward completing the data conversion efforts which Archives began eight years ago. The final product will provide us over 6,700 MARC records and a similar number of EAD files describing our collections. In addition, we will be providing enhanced access tools and the ability to quickly and easily add links to digitized items into our holdings records. The project will provide a model which can be applied in other University of Illinois units as well as other institutions. The University Archives and those associated with the Library's emerging digital library will develop technical expertise which can be applied to other digital projects and statewide or regional services to be developed at a future date.