Developing a digital archive with limited resources

Greg Sennema

The author

Greg Sennema is Digital Resources and Reference/Instruction Librarian, Hekman Library, Calvin College and Calvin Theological Seminary, Grand Rapids, Michigan, USA.

Keywords

Academic libraries, Digital libraries, Internet

Abstract

This article describes how a small liberal arts college library created a digital library using existing staff and financial resources. The Hekman Digital Archive (HDA) was created using a digital media archive module of the library's integrated library system, and is maintained by one library staff member with the help of numerous student assistants.

Electronic access

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1065-075X.htm

Staff at the Hekman Library first considered developing a digital archive while investigating options to upgrade the library's integrated library system (ILS) in the spring of 2000. Primarily serving the students, staff and faculty of Calvin College and Calvin Theological Seminary in Grand Rapids, Michigan, the Hekman Library provides access to over 1.6 million items including approximately 175,000 e-books and over 10,000 e-journals. Along with the myriad issues library staff needed to consider in a new ILS, staff were also mindful that individuals within Calvin's academic community were interested in storing digital files that could be used for classroom instruction. With this in mind, and after reviewing proposals from three ILS vendors, the library purchased an ILS from Sirsi which included Hyperion, a module for cataloging, storing and making digital files accessible on the web. This article describes the events, processes, challenges and decisions that brought the Hekman Digital Archive (HDA) into existence, and turned it into the valuable resource that it is today [1].

Digital media archive: getting started

The decision to purchase a new ILS was made in the spring of 2000, and in the fall of that year Sirsi provided on-site training to the library's 17 FTE staff. In addition to providing training for the catalog module (Sirsi Unicorn), Sirsi also scheduled specific training sessions for creating a digital library using the digital media archive module Hyperion. At the time of this training, library staff had little prior knowledge of what a digital archive was supposed to be, including the basic definition of "digital archive," and how it might differ from or resemble what library literature commonly referred to as a "digital library". Furthermore, no significant money or additional staff could be devoted to the digital archive beyond the cost of the Hyperion software and a new server.

Other stakeholders – the library director, the library systems programmer and the curator of the college's archives – were invited to participate in the two day on-site training. In addition to providing guidance on how to use Hyperion, the Sirsi representative helped participants develop a "pilot project" by reviewing issues such as staffing, handling of materials, scanning specifications, digital masters and derivative images, Dublin Core metadata standards, quality control and delivering the digital files to the web. For a pilot project, library staff identified 700 historical images from the college and seminary archives that would be digitized in

conjunction with Calvin's 125th anniversary celebrations in May of 2001 [2]. Using a pilot project as the basis for training proved helpful in formulating the basic workflow that would be employed for the digital archive. The ultimate goal was to develop foundational infrastructure, rather than to recreate the infrastructure each time a new digital collection was created. As such, library staff wanted to mimic the procedures – purchasing, processing, cataloging, shelving, etc. – that were already in place in the library's acquisitions and cataloging departments.

The digital archive's "technical services" area was created immediately after training. Both a flatbed and slide scanner were purchased, and two student assistants were hired to complete the scanning and basic metadata for each image. Scanning for the digital master was completed according to specifications outlined in the literature at that time (e.g., Kenney and Rieger, 2000). During the time it took for each image to be scanned, the student assistants recorded the images, descriptive information onto a spreadsheet using metadata rules defined by the digital archive's metadata elements.

The plan to create the 125th Anniversary Archives using Hyperion was not realized because of the time needed to properly configure the library's new ILS, and the author's need for additional training. Library staff decided to create a temporary Web-accessible digital collection, and the library systems programmer swiftly produced a searchable Web portal which displayed records that included metadata, thumbnails, and various resolution sizes. An online administrative component allowed library staff to add, edit, and delete the images and the metadata in this temporary portal.

After the 125th Anniversary Archive Web site was completed, the author focused on creating the infrastructure for a permanent digital archive. The library had a server to store the digital files, client-based software for managing records, and a web-accessible portal. The then current version of the Hyperion software, however, was difficult to use. Using it to upload digital files, move and manipulate menu hierarchies, and enter and edit metadata could only be completed with deliberate effort by a metadata librarian. Calvin, however, did not have a metadata librarian, and it was unlikely that already overworked library staff and student assistants – working sporadically on digital archive tasks – would be able to master Hyperion without constant facilitation. To overcome this considerable obstacle, the author envisioned that the work could be shared with other library staff and student assistants using simple web-accessible forms.

Hekman Digital Archive (HDA)

The challenge was to create a digital archive without the oversight of a full-time staff person, while at the same time creating an efficient workflow that ensured high-quality digital files and metadata, and a timely turnaround on processing this data. Another consideration was how to display the permanent digital archive to the Calvin community. Library staff decided to make the digital media archive resemble the library's online catalog, using similar layout and graphical features, and to call the resource the Hekman Digital Archive (HDA) instead of using Sirsi's brand name Hyperion. The HDA would be comprised of digital collections arranged according to a hierarchical menu system (as defined by Hyperion) with searchable metadata. The first official HDA collection after the pilot project was entitled "Calviniana", comprised of digitized files of images and medals (see Figure 1) depicting John Calvin collected by the H.H. Meeter Center for Calvin Studies [3]. The creation of this collection was helpful since it offered the opportunity to see Hyperion in action, and to further understand its capabilities and limitations. It also provided staff with additional scanning experience, and in particular, experience in digitizing three-dimensional objects.

Hekman Digital Archive Queue (HDAQ)

The amount of time spent on creating Calviniana, particularly the effort expended using Hyperion for this collection, convinced the author to find an alternative workflow for the digital archive. The library systems programmer suggested an alternative based on the 125th Anniversary Archives. A modified version of the administrative component of the Anniversary Archives could be used as the environment (or "software") that library staff and student assistants would use to create the digital archive content. As with the Anniversary Archives, student assistants could be employed to complete the digitization process and simultaneously to record the basic metadata for each file. This online administrative module was named HDAQ, with the "Q" referring to the queue of records that are waiting to be uploaded into the HDA by library staff after they are created in the HDAQ by student assistants.

After digitizing the images, the student assistant would log into HDAQ with a username that predetermined particular collection variables. Clicking "Add a new entry" (see Figure 2), the student processes the image. In HDAQ, this means that a piece of software named ImageMagick is

Figure 1 The "medals" collected by the Meeter Center are comprised of coins and other commemorative medals. The medals were digitized using a flatbed scanner. In addition to the thumbnails and metadata, this figure shows the "medium resolution" image in a pop-up window
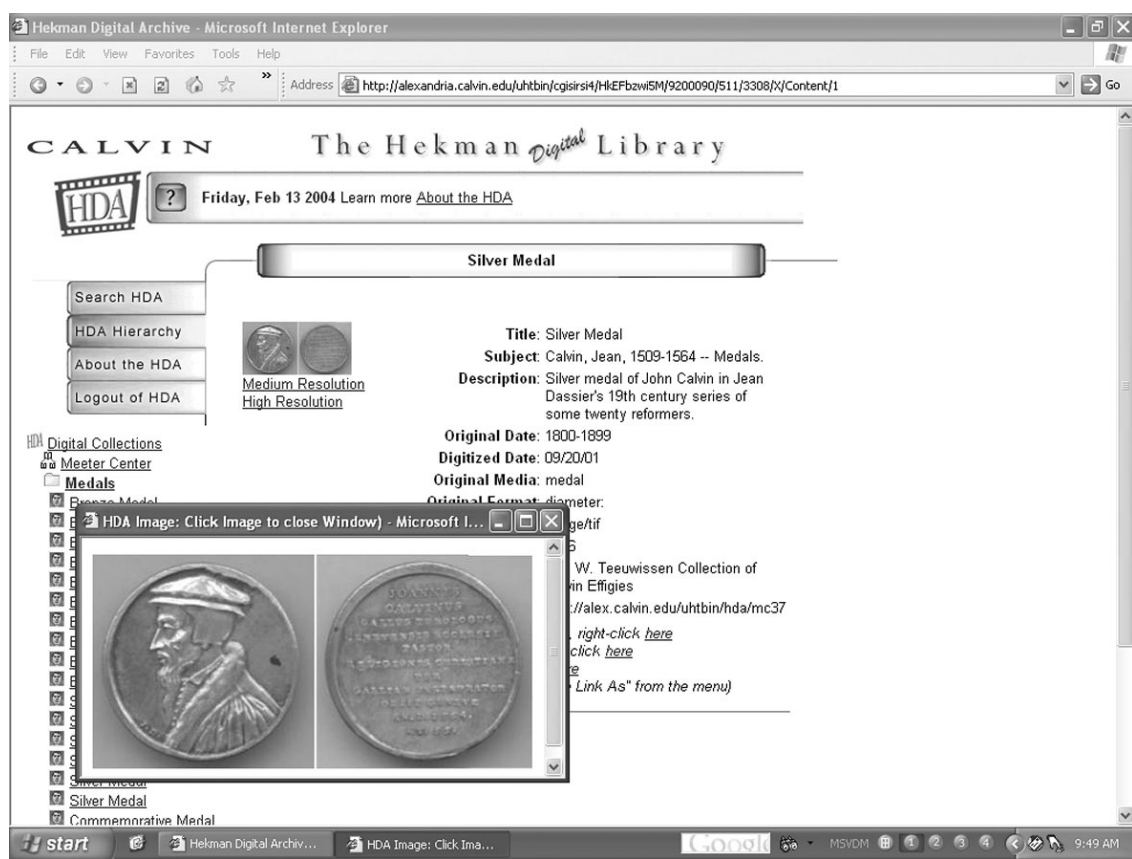
utilized to create various lower-resolution JPEG images (including a thumbnail) from the original higher-resolution TIFF image. During the several minutes that this process takes to complete on the server, the student enters metadata onto an online form. By the time the metadata has been entered, the image processing has been completed, and the next record can be created.

In addition to creating HDA records, HDAQ is also used to manage characteristics of each collection. For example, HDA records are placed within a hierarchical menu. During the initial Hyperion training with Sirsi, library staff decided that the hierarchy would contain four top-level menu items: the Hekman Library, Heritage Hall, the Meeter Center, and Academic Departments. Beneath these four menu items, submenus could be added according to the type of collection. In the case of the Meeter Center, the Calviniana collection was placed immediately below the main menu item. Beneath the Academic Department menu item the names of Calvin's departments would be listed, and beneath these the names of various collections created within the department. HDAQ logins could be created that would force all

created records into a particular menu item. This feature would later become important when additional student assistants were hired to create records for an increasing number of collections. Finally, HDAQ is the location in which staff can access the queue of records created by student assistants in order to review the quality of the image and the metadata, and to add Library of Congress (LC) subject headings before each record is uploaded to its final location in the HDA. After being loaded into the HDA, the records are no longer accessible from the HDAQ, and subsequent changes to the images or the metadata are completed using Hyperion.

HDA records rely on two file naming schemes, one for the original digital file, and the other for the HDA record. Since each HDA image is part of a digital collection, it is assigned a letter code abbreviation (based on the collection's name) followed by sequential numbers. After being entered into HDAQ, the image number information is placed in the "Identifier" metadata field, while the HDAQ record (the image along with the metadata) is assigned an auto-generated number based on the department to which the

Figure 2 HDAQ's main screen
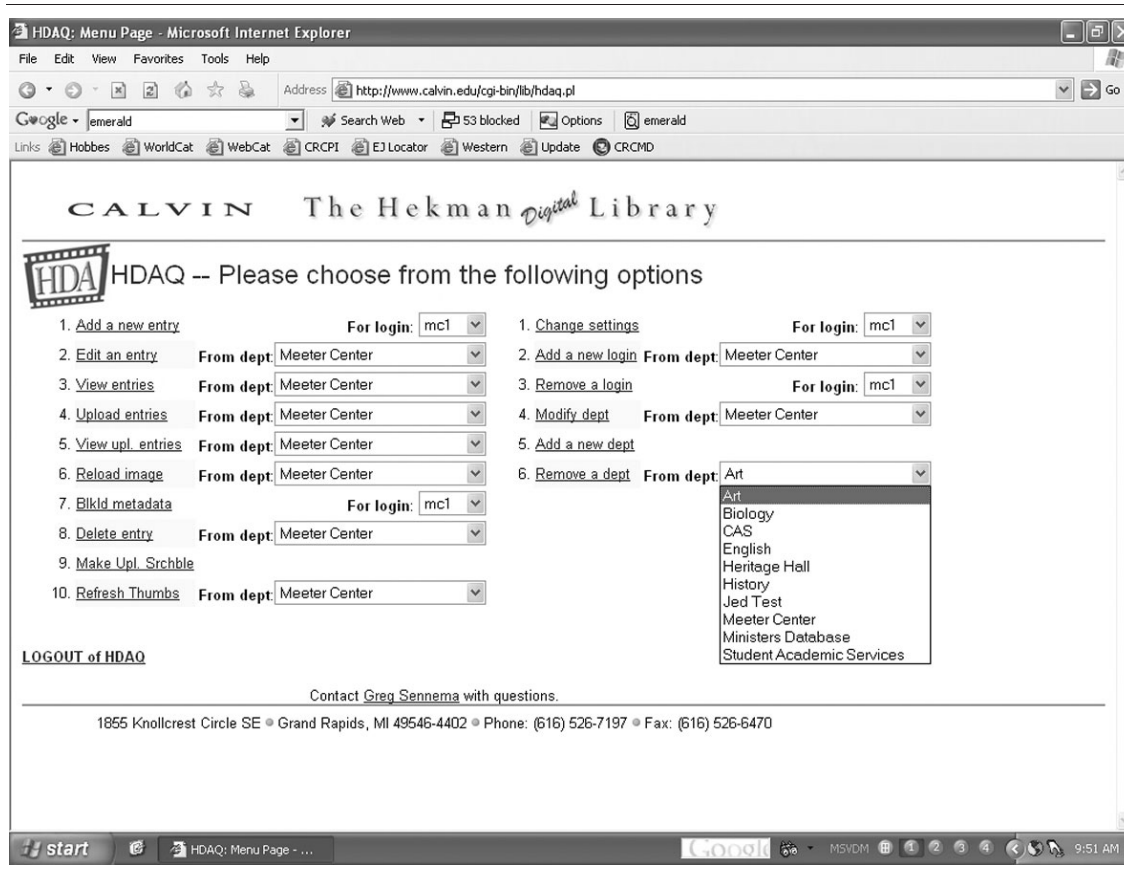
digital collection belongs. For example, the final image in a collection of theater history slides might be "th1672", and the first number of a subsequent collection of costume history slides be assigned as "ch1", but their corresponding HDA identifiers would be auto-generated based on their entry into the system.

Adding HDA collections

In May 2001, with one HDA collection completed (the Anniversary Archives would be incorporated into the HDA the following year), library staff were approached by a professor from the college's Communication Arts and Science (CAS) department who requested that the library create a digital collection of slides covering the history of theater. Library staff were pleased with this request, since the professor offered the services of a student assistant to complete the scanning and basic metadata. Unlike the Anniversary Archives or Calviniana images, the primary purpose of the digital slides in the Theater History Digital Archive Collection was for classroom instruction, and in this case, the image quality did not have to meet the high-resolution specifications followed

for the previous collections. Although copyright was not an issue for this collection, it is accessible only to members of the Calvin community. Hyperion allows for various levels of access to individual records or whole menu items, including an "Extended Public" level which makes a resource freely accessible on the web, and a "Public" level which requires users to log in using ID and password information stored on the college's authentication database.

The Theater History collection offered library staff the first occasion to use HDAQ. The student assistant was trained to scan the slides and to record metadata from information printed on the slide frame, with additional information coming from monographs and other online sources. After the scanning and data gathering was complete, the student assistant was shown how to use HDAQ. His HDAQ login determined that all the records he created would be automatically assigned an access level of "Public", and that they would be placed in the "Academic Department/CAS/Theater History" position of the hierarchical menus. Further menu divisions, such as "/Medieval/Western European/Miracle Plays" were specified when the metadata was entered into HDAQ. As the student assistant

continued to add records in HDAQ, library staff reviewed the work, added LC subject headings, and "uploaded" each record. Upon completion of this collection, library staff noted that HDAQ proved adequate to accept the image files and metadata, as well as to synchronize this data to its final home in the HDA. Library staff also realized that it was not possible to create one workflow that would effectively deal with each digital collection. The images in the Theater History collection, for example, were scanned at a lower resolution since the professor was not interested in preservation as much as in the immediate access of the data by students and faculty in the classroom. In addition to image quality, other questions had to be addressed for each collection. For example, would student assistants be hired by the library or by a college department? Would the same person complete the scanning and the metadata? Where would the collection be placed within the hierarchy in relation to other collections from the same department? Would the collection be available to the public, or restricted to the Calvin community?

HDA: defining a purpose

The library was also approached by the college's Public Relations staff, who proposed storing their digital files on the HDA. PR staff wanted to manage their digital images in an intuitive system – more than just the simple network directory they were using – that allowed each image to be indexed and arranged within directories for easy retrieval and manipulation for their PR purposes. After reviewing their proposal, the library declined their request. Library staff suggested that they consider using photo-management software, and pointed out that they were asking us to use the HDA in a manner analogous to using an online catalog to manage their office documents and publications. The author took this opportunity to refine HDA documentation by clarifying the HDA's purpose as a tool to complement classroom instruction, and to describe what types of digital collections would be considered for inclusion [4].

To expand the number and types of HDA collections that could be used for in-class instruction, library staff approached professors from Calvin's history and biology departments who were known to incorporate visual aids into their instruction. None of the professors was able to commit time to scan or record metadata, so library staff offered to complete the majority of this work, requesting only that the professors commit to adding metadata to a few fields, particularly title and description. All metadata was recorded onto a spreadsheet containing column labels

(metadata fields) and rows (records) allowing recurrent data to be easily copied and pasted. Once complete, the student assistant would use the spreadsheet to enter each record into HDAQ. By this time (June 2002), the library systems programmer had incorporated the Anniversary Archive records into the HDA using a script that converted records from HDAQ's SQL-queried database tables into files acceptable to a batch-load utility in Hyperion. Library staff soon realized that this script could be used to batch load spreadsheet data directly into the HDA through HDAQ. This batch load utility resulted in a quick turnaround time from collection conception to realization. It also allowed library staff to reload collections if revisions were necessary [5].

By the summer of 2003, library staff had enough experience in creating HDA collections to entertain a request from Calvin's Art department to digitize their collection of over 10,000 slides. With so many slides, library staff realized that in order to incorporate this number of images into the HDA in a timely manner, student assistants would need to be involved in reviewing metadata and completing subject analysis. Using a combination of Library of Congress Subject Headings volumes, OCLC WorldCat, the Getty online Provenance Index, and other online searches, the student assistant was able to efficiently and effectively update and upload over 1,500 HDAQ records within two months.

Conclusion

The success of the HDA is difficult to measure, especially when compared to some of the impressive digital collections being made available on the web, collections with innumerable records and polished user interfaces. On the other hand, library staff were able to move from a vague concept to a digital archive of over 5,000 records within three years. Furthermore, this work was accomplished using less than 10 percent of the author's time, less than 10 percent of a part-time systems programmer's time, and perhaps 20 percent of student assistants' time. With the current level of staff time devoted to HDA collections, it is difficult to imagine a future that is much different than merely adding more records, collection by collection, department by department. However, increasing professional staff time devoted to the HDA would significantly enhance the usefulness of the HDA, opening up the possibility of accommodating additional formats such as audio, visual and full text, and recovering control over metadata and subject standards. Additional staff would also allow the

library to incorporate more diverse and larger collections at one time, or to quickly create collections for special events and exhibitions. Finally, library staff would be able to work closely with the college's Information Technology (IT) and Audio-Visual departments, providing seamless access between the library's archival collections and the rest of the technological initiatives throughout Calvin's campus, and through merging areas of expertise represented by all these departments.

Notes

1 Visit the HDA at www.calvin.edu/library/tutohelp/hda/index.stm
2 The Heritage Hall Archives serve Calvin College and Calvin Theological Seminary as well as the Christian Reformed Church, the denomination to which the college and seminary belong.
3 The H.H. Meeter Center for Calvin Studies: see www.calvin.edu/meeter for more information.
4 See www.calvin.edu/library/tutohelp/hda/purpose.stm
5 After viewing the "Plant Taxonomy" digital collection in the HDA, the biology professor pointed out the protocol of displaying plant species in italics. Adding italics, or other types of formatting, is not offered as short-cut options in Hyperion, and would require that the title field of each record incorporate the html code for italics (e.g. $<$i$>$Osmunda regalis$<$/i$>$). The batch load utility made it possible to quickly reload the data with the proper html codes.

Reference

Kenney, A. and Rieger, O. (2000), *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, Research Libraries Group, Mountain View, CA.