Harvesting the Low-hanging Fruit: World Wide Web Access to a Collection of MIT Theses

Larry Stone, MIT Information Systems

Bill Comstock, Harvard College Library[1]

Keith Glavash, MIT Libraries

Abstract

Almost by accident, we have created a digital library mounting a growing retrospective collection of 4,000 MIT theses. By taking advantage of existing procedures in the MIT Libraries we are accumulating scanned images of selected theses (the low-hanging fruit), and with some freely-available software and a little development effort we put them online. Even without any active promotion on our part, it has been discovered and used by scholars all over the world, demonstrating its value as a source of knowledge. In this paper we describe this digital library and trace its evolution.

A User's View

The *Digital Library of MIT Theses* (DLT) at `http://theses.mit.edu` is a fairly typical digital library website: the user chooses documents by issuing a search or browsing the index, and then views them in a separate reading mode. The DLT's look and feel will be familiar to anyone who has used the *Networked Computer Science Technical Reference Library* (NCSTRL)[1] since it is built on the same *Dienst[2]* software. It emits basic HTML that works equally well with all Web browsers, although an image-capable browser is required to see the pages.

Its search capability is fairly primitive as yet, limited by both Dienst and the metadata we have available. The *author* and *title* of all documents are indexed for searching, but only some of the theses have *subject* keywords indexed. We are in the process of adding the full text of abstracts, which will provide a much more effective index for subject searches.

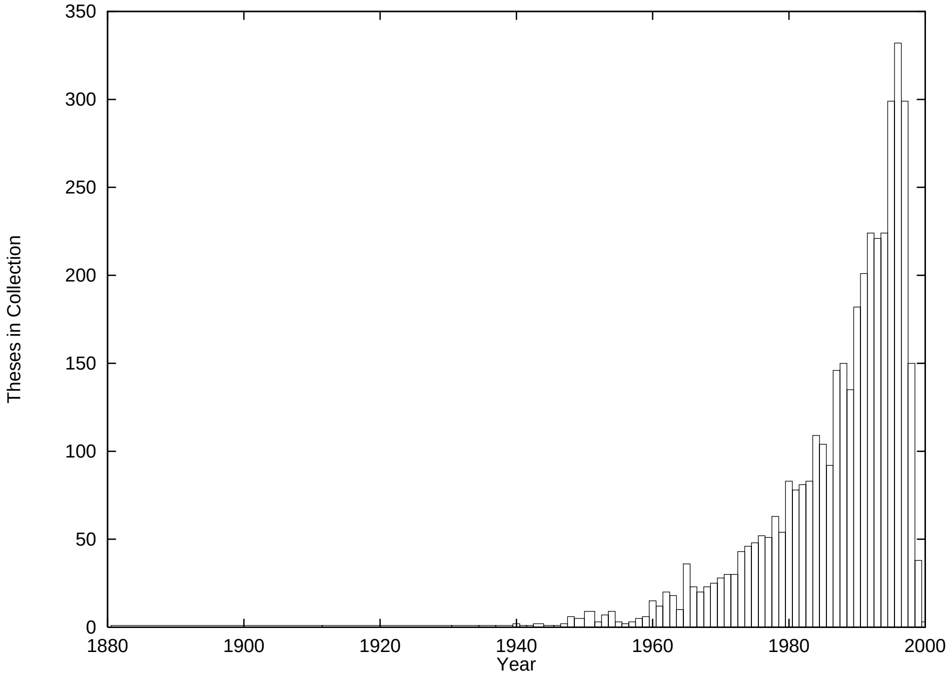
The Dienst software can also list groups of documents sorted by date or by author. This is convenient for patrons unsure of how to spell an author's name.

The pages of search results and "browse" lists present each thesis as a link leading to its "summary" page, which lists its viewing options. All theses can be viewed one page at a time as grayscale page images, or as arrays of "thumbnail" images giving an overview of 24 pages in each screen. The thumbnails are an effective navigational aid, since they make it easy to find major features like the table of contents, chapter divisions, and key figures.

[1] Formerly of the MIT Libraries.

The summary page also offers an option to purchase the thesis from Document Services as either a bound paper copy, or electronically-delivered, high-resolution printable PDF. The latter option has been popular with customers in countries where physical delivery of books is expensive and unreliable. The "purchase" link leads to a Web form with all of the metadata identifying the thesis filled in. Future plans include secure credit-card ordering and faster, simpler online delivery than the current FTP process.

The DLT currently houses almost 4,000 theses dating from as far back as 1896. Over half of them are from the last 10 years, however.   Since (as we explain later) theses are added when customers buy printed or electronic copies from Document Services, these are mostly the theses in demand by paying customers:



Windfall: Discovery and Acceptance

Because it was initially intended as an experimental service, at first we did nothing to promote the DLT or even announce its existence. We linked it to the MIT Libraries website but forbid search engines from indexing any of the DLT. Nonetheless, many researchers, libraries, and World Wide Web explorers made their way there. *Yahoo!* found it, and lists it under "Reference > Libraries > Digital Libraries > Electronic Theses and Dissertations (ETDs)." It has also been linked by other research libraries, the Internet Scout project, and even *USA Today's* "hot sites".

We haven't done much to track usage, aside from informally monitoring summaries of statistics collected by the web server and digital library software. We have no detailed long-term data since we delete all access logs after a month, keeping them only long enough to generate summaries.   This is done to conform to the common library practice of discarding all records of the specific materials used by a patron.

On average, users request about 2,000 page images from the site each day. Log summaries also show that at least 20 theses appear to be viewed, in their entirety, every day. There are over 100 referrals per day from sites such as Yahoo,

The DLT site has not generated many email responses from patrons, although the contact address, *etheses-help@mit.edu*, is mentioned on every page. It typically gets only a few messages a week, and most are confused patrons asking us to do searches and answer reference questions. Only one message in the past year even mentioned a bug in the system, although there have been other bugs and outages.

Developing the Digital Library of Theses

Planting the Seed

It all began with a project aimed at capturing the scanned images produced incidentally in the Document Services department of the MIT Libraries. Normally, Document Services technicians scan the archived microform of the thesis to produce 600dpi bitonal images which are then printed on a laser printer. It was done this way because the microform is easier to handle and can be scanned automatically; the scanned images were deleted after each job.

Our initial plan was to set up a private repository to keep the scanned images online. Then, the next time a scanned thesis is requested, the image files are just printed again, skipping the retrieval and digitizing steps. Since about half of the theses requested have already been printed recently, this would save a considerable amount of work.

The thesis images were initially stored on a Unix-based server in a directory named with the thesis' *OCLC number* (a unique number in its catalog entry, assigned by the Online Computer Library Center). A thesis' directory contained single-image TIFF files, one for each page scanned.

We constructed a simple Web interface through which Document Services technicians could upload and download theses, automatically sending the entire directory of image files in one transaction. It turned out to be more effective to collect newly-scanned theses automatically, however, so the upload function is rarely used. The Web interface to the repository is secured, restricted to Document Services personnel identified by their X.509 Web certificates.

The automatic upload mechanism takes advantage of the fact that as they are scanned, image files are actually written to a another Unix workstation, in Document Services, via NFS. They are placed in a directory named by the thesis' OCLC number. Every night, a daemon on that workstation looks for completed theses and uploads them to the repository server through an anonymous-FTP drop-box. This two-tiered approach is required because NFS is not a secure protocol and may not be run on exposed, public server machines at MIT.

Since the capture of theses began in January of 1998, the Document Services technician printing a thesis first checks if it is online by giving the OCLC number to the repository's Web interface. The thesis is downloaded if it is available, and then it is sent to the printer. This simple system met the original goal of expediting the printing workflow for theses that have already been scanned, but the online collection was still not available to the public. It was also quietly building an online collection of scanned documents that could serve as a testbed for another project, such as a digital library.

Serving the Fruit

To turn this repository into a real digital library, we needed metadata, and software to present all of it to users. Fortunately, both of these were, like low-hanging fruit, almost within reach.

Acquiring metadata

Fortunately, the key to getting metadata was already available in the OCLC number used to name each thesis directory. In *Barton*, the MIT Libraries' online catalog, there is a one-to-one mapping between all holdings for a thesis and one bibliographic record indexed by an OCLC number, so each number precisely identifies a thesis. This was also a stroke of luck for us, since some institutions have different OCLC numbers for each archived copy of a thesis, e.g. the microform and the paper. With a simple Z39.50 client that retrieves the *MARC* (MAchine-Readable Cataloging) record matching an OCLC number lets, we could acquire authoritative metadata automatically.

The remaining task was perhaps the most difficult one. We had to extract the relevant metadata elements from each MARC record—and since the thesis collection spans many decades, there is some variation in the cataloging conventions used to create them. For example, the academic department and degree are represented in the 502 field (thesis notes), but there are at least three different conventions for arranging and punctuating the text of that field.

Once extracted, the metadata are written out in a special simplified format used by Dienst. We chose to translate the metadata to this format rather than adapt Dienst to read MARC for pragmatic reasons: The code to interpret MARC records is all in one application and is easier to maintain. We anticipated the need for extensive debugging of the translator, but have been pleasantly surprised by its robustness. Using Dienst's native metadata format also minimizes the modifications made to the Dienst code.

Dienst's native metadata format is the "BIB" record, defined by the Internet RFC 1807, *A Format for Bibliographic Records*[3]. This (as well as Dienst itself) was developed for the *CS-TR* (Computer Science Technical Report)[4] project. It is a simple text-based format consisting of a sequence of named fields for elements like TITLE, AUTHOR, KEYWORD, etc. To accommodate metadata elements peculiar to theses which did not fit into any of the RFC 1807 fields, we added three extension fields, to make those metadata available to Dienst if we want to build a searchable index out of them later.

X_DEGREE	Degree granted for thesis, or “none”.
X_DEPARTMENT	Academic department(s).
X_SUPERVISOR	Student’s primary supervisor.

The process of translating metadata also gives us a measure of quality-control. It detects an invalid OCLC number because the catalog access fails. If the number is valid but refers to the wrong document, a randomly-selected catalog entry is probably not a thesis so the translator warns about that when the 502 field unique to theses is missing.

Digital Library Software

We chose the Dienst software because it could do everything we wanted in the way of managing and presenting documents, and because we were already quite familiar with it from maintaining MIT’s NCSTRL site. We deployed the newest version then available, Dienst 4.1.9, released in 1997. Designed for earlier HTML standards and practices, it looks primitive now, but at least it is compatible with older browsers. Performance might be a problem, since Dienst is written in interpreted Perl, but a powerful CPU on the server makes that less of an issue.

We configured Dienst to offer only low-resolution page images, suitable for viewing on desktop display screens. Document Services counts on sales of printed and bound theses to recover the cost of the initial scanning, so distributing free high-resolution images would cannibalize their market.

There was one more conversion problem: The pages are stored as high-resolution (600dpi) bitonal, 90-degrees rotated TIFF images for Document Services printing requirements. Since the popular Web browsers display images exactly received without any options of scaling and rotating, we had to send pages that were suitable for a desktop display. After some experimentation we found 100dpi resolution and 32 levels (5 bits) of grayscale to be the best compromise between small image size and good legibility. We chose CompuServe GIF as the commonly accepted image format with the most appropriate compression and color representation characteristics.

The image translation is done on the fly by a custom C program that rotates, reduces and writes out a full 8.5” x 11” page in about half a second. Since keeping both kinds of images would take twice the disk space, we decided to have the low-resolution ones produced dynamically. The extra half-second lag is not very noticeable when added to the transmission time of a 50-100 Kb page image for most users.

Finally, Dienst needed a naming scheme (called “Document IDs” for the theses. We chose not to use the OCLC number that names the directory, mainly because this number is not under our control and its entire nature might someday change. Although the DLT was set up as a short-term experiment, we tried to prepare for problems we could envision even a decade or two into the future, to minimize the changes required to make it permanent. We created a new naming convention to identify theses in the DLT: The 4-digit year of publication, followed by a dash and a decimal number to make it unique. The unique number was incremented from one for each year, yielding doc-id's like `1996-1, 1997-1, 1997-2,` etc. Since we were issuing these numbers there would never be a reason to change them.

Hardware requirements

The server is a Sun UltraSPARC workstation running Solaris 2.6. We chose the Apache web server because it is efficient, reliable and easy to configure securely. Apache 1.3 requires a small modification to get along with Dienst. We are using Sun's Solstice DiskSuite software to manage the storage (a hardware RAID), since it allows filesystems to be dynamically grown. This lets us keep all of the theses in one filesystem even as the collection is growing at the rate of 20Gb a year; by the time we reach the 1 terabyte filesystem limit, the operating system technology will likely have been improved to extend it.

Theses vary greatly in size but on average they are about 200 pages and occupy 15 Mb of disk space each. We presently have about 60 Gb of theses online.

Refinements: Bugs in the fruit.

We first brought up the digital library of theses in September, 1998 after about a person-month of development effort. What appeared at first to be low-hanging fruit still needed some ripening, however. The initial service had many gaps and rough edges which have since been filled in with the following developments.

- A new link on the summary page of each document leads to a filled-in order form, which patrons can use to purchase a either a printed copy or an electronically-delivered high-resolution printable PDF file from Document Services.
- The Dienst software was configured to produce "thumbnail" navigation views (automatically, of course) for each thesis. The thumbnails, each of which links to a full-size page, are surprisingly useful. They make it easy to locate major structural features like the table of contents, and important graphics. Thumbnails compensate somewhat for the lack of structural metadata.
- We gradually improved the collection and handling of metadata. Subject keywords in the MARC records are added to the Dienst "abstract" index for searching. We fixed a character-set problem that we chose to finesse at first: The MARC-21 character set is superficially similar to the Web standard character set, ISO 8859-1, but the non-ASCII characters are quite different. Some author names were not indexed correctly and searches would fail until the MARC translator was fixed. It now converts MARC-21 characters accurately whenever possible and renders them as unaccented text when there is no equivalent in the ISO 8859-1 character set.
- Some simple changes in the user interface had a profound effect. We used to get frequent email from patrons asking why a thesis was not in the collection, when the author was known to have written a thesis at MIT—the bold declarations on the home page that this is a partial collection were not being noticed. After we improved the wording

on the front page and added notices to the page returned for a failed search, those questions became more rare.

- Incorrect image formats and, occasionally, corrupt files and mistyped OCLC numbers in newly-scanned theses caused annoying failures that required handwork “behind the scenes” to unravel.
- When a thesis with incorrect image formats or corrupt images was uploaded and installed in the digital library, human attention was required to unravel the mess. By moving some quality checks downstream—even doing them on the NFS server in Document Services before the theses are uploaded to the digital library, we minimize the handwork required on the server. The Document Services technicians doing the scanning can diagnose and correct problems with image formats more easily before the thesis is uploaded.
- Much additional development effort expended on the DLT has been directed at increasing the degree of automation in the system so less human intervention is needed. Periodic daemons monitor Dienst and restart it if needed, and check the consistency of the digital library’s contents.
- On the strength of this collection and a pilot project in electronic thesis submission, MIT has joined the *Networked Digital Library of Theses and Dissertations* (NDLTD). Server statistics show that patrons are finding our collection from the `www.theses.org` site, and perhaps we will be able to join the NDLTD’s federated search system to make our content accessible to more patrons.
- Theses are now given handles in the *Handle System*[6] as they are installed. The Handle System assigns persistent names to networked resources, in the manner of URNs. Since our thesis handles are currently in the namespace of an experimental local installation of the system, we cannot yet cite them as truly persistent names, but the DLT is ready to adopt a permanent Handle System service immediately when we have it.

Future work

Quality control of the scanned page images is still not perfect, and because the online theses are mostly generated as a side effect of printing paper copies, scanning problems are often not corrected because they do not need to be: If a couple of pages scanned poorly or were unclear in the microform, the technician “tips in” pages copied from the archived paper version of the thesis. These repairs are not propagated to the online copy, although that can someday be done with better software support. Now that Document Services is offering printable theses in electronic form, we have more motivation to ensure that the online copy is perfect.

We would like the MIT Libraries catalog, Barton, to list the online theses as holdings. Then the Web interface to Barton would show them as HTML links. This is technically feasible in its GEAC ADVANCE software, but the resources have not been available to implement it.

The digital library might also use Barton as its search engine and index, once it is possible to distinguish the DLT contents (i.e. those theses with online copies) in queries. Dienst was designed to let an alternative search engine "plug in" to take over from its internal catalog, so the implementation ought to be straightforward. Searching in Barton would also let us search all MIT theses and show patrons relevant theses which are not yet online.

A project now underway will give us the full text of the abstract and table of contents for each thesis. We do not have the resources for OCR, but an outside agency is doing the OCR for their own needs and then sharing it with us. Though far from perfect, the accuracy should be good enough to use the abstracts for "subject" searches and perhaps even to present to the user. The only "subject" index we have now is taken from MARC records, but since not all thesis records even have subject keywords its coverage is uneven.

The format chosen for document identifiers turns out to have been a mistake. We named documents with the year of publication followed by a dash and then a unique number (e.g. `1996-13`), but the error was in basing the identifier on the date. Because of that, we have to (inconveniently) get the metadata for a thesis before assigning it a document ID, and that ID now preserves a metadata item which might, in the future, be revised. This has already happened within our relatively small collection; The year of a thesis was changed when a cataloging mistake was corrected. In general, it is best when the digital library document identifier is completely unrelated to the content of the document, as are catalog reference numbers like the OCLC number. Then the identifier can be assigned blindly, and users are not mislead into trying to divine the contents of a document from the identifier (or guessing identifiers). We may yet change to a new system of opaque document identifiers before putting them into truly permanent Handle System handles.

Conclusions

We have discovered that with a small initial effort we can build a usable digital library out of a pile of raw digitized documents. With material as original and unique as MIT theses, even a partial collection of imperfect quality is a valuable enough resource to attract researchers from all over the world, especially when it is freely available and convenient to access.

Automation was an essential component of our success, and a key ingredient of the low-hanging-fruit model. Human attention is expensive, so documents have to be acquired without disrupting the normal workflow in Document Services, or requiring much effort from programmers in Information Systems. The digital library website runs and grows with virtually no maintenance, and only a little time spent on operations tasks.

Interestingly, Document Services has not noticed much change in the rate of orders for printed copies of theses. Perhaps the customers who now settle for Web-viewable theses are being replaced by new print customers who were first attracted by the website. More people in total are now reading the theses, which fulfills MIT's mission to make its intellectual products available to the world.

The digital library of theses has generated enough excitement to take it beyond the stage of a tentative experiment, to likely become permanent fixture at MIT.

Acknowledgements

The project that developed into the DLT was made possible by computing hardware and staff resources allocated by Vijay Kumar, Assistant Provost and Director of Academic Computing, for further experiments with digital libraries. It was also built on the groundwork and tradition of Information Systems / Libraries collaborations developed by William D. Cattey through the MIT-PIDS project and other endeavors.

References

[1],Networked Computer Science Technical Reference Library (NCSTRL), see `http://www.ncstrl.org`/.

[2]*Dienst: Implementation Reference Manual*, Carl Lagoze, Erin Shaw, James R. Davis and Dean B. Krafft, handle: `ncstrl.cornell/TR95-1514`

[3]Internet RFC 1807, *A Format for Bibliographic Records*, Danny Cohen and Rebecca Lasher.

[4]Computer Science Technical Report (CS-TR) project, see `http://www.cnri.reston.va.us/home/cstr.html`

[5]Networked Digital Library of Theses and Dissertations, see `shttp://www.ndltd.org`

[6]CNRI's Handle System, see `http://www.handle.net/introduction.html` for details. Also see the CNRI home page: `http://www.cnri.reston.va.us/`

[7]MIT-PIDS—*The MIT Page Image Delivery System Architecture Manual*, Bill Cattey, see `ftp://athena-dist.mit.edu/pub/elib/arch_pids.PS`