The LLMC-Digital Newsletter
Issue No. 11: December 2004

Going Agog Over Google

The big news for now, of course, is that Google is thinking big. Unless you honeymooned on Mars last week, you've probably heard something of Google's ambitious plans to digitize all the books in several major libraries, as much as 15-million volumes in total. This, not so incidentally, includes some 7-million volumes from LLMC's digital-library partner, the University of Michigan (UofM).

The press, of course, chose to play this as the setting of the sun on brick-and-mortar libraries and the dawning of a tomorrow when "everything will be on the web." Of course, not everyone is so sanguine. For example, in an e-memo widely circulated among law librarians, Dale Askey of Yale's Sterling Library critiques Google's plans to scan the 7-million volumes at the UofM over six years as follows: "Assuming that one could run this project 24/7 for all six years at full tilt, one would need to digitize 133 books per hour for the entire period. Taking a conservative page-count of 150 per volume, that's a whopping 19,950 pages per hour…. Can you imagine the cost of doing such industrial digitization work inside the US?" **(endnote 1)**

Dale from Yale may be on firmer ground when he points to potential trouble for Google related to encoding and OCRing the texts. Our experience at LLMC is that these two functions add substantially to post-scanning costs and grief. To put things in perspective, LLMC has a goal of digitizing 100-million pages in 10 years. Google is aiming at about 1.05-billion pages in just the six-year span planned for the UofM scanning. That's ca. 125-million pages per year! Given the sheer gigantism of this project, snafus will likely evolve. The true wonder in this tale will be if something wonderfully weird does not go wrong.

Nevertheless, all of that should worry only Google and its new stockholders. The rest of us should rejoice in the prospect that, at little cost to the public sector, many millions of almost-lost texts, now moldering in near-inaccessibility in off-site storage, may become easily available and digitally browsable for the whole world. Most of these books have no moneyed constituency to pay to let them see the light of day. They are orphans for whom a Daddy Warbucks apparently has magically appeared. We can only wish both Google and the UofM well. As to whether the project ever achieves its lofty goals in full, as one sage Michigan librarian put it "off the record":

Page 2 begins here

"We've always figured that if we come out of this with, say, a million volumes, we're still way ahead of where we would have been on our own."

Of course, given our parochial interests, one fly in the ointment could be how

Google's project might affect *LLMC-Digital.* After all, some may ask, if Google will do all of this stuff anyway, and for free, why should we pay to do it also? A full answer to that question will likely take a decade of experience to work out, but some preliminary observations are possible even now. **(endnote 2)**

Access and usability for large quantities of digitized data will mirror the investment in enhancement during its aggregation. The level of investment may reflect desire, but will rely on ability. In the latter category, certain affluent professions like medicine, engineering and law are going to have much more ability to invest than will, say, the Fraternal Order of English Majors. This may not be fair, but, as JFK once observed, life is not fair. We apparently are evolving into a two-tier digital world where some classes of heritage literature will be massaged to a higher level of accessibility and utility when they go digital than will others.

Googling is great, but, when it comes to navigating specific bodies of specialized literature, it still represents the "plain vanilla" approach. For an example close to home, even if all of the relevant legal data were in both systems, no one seriously expects Google searches to replace Lexis or Westlaw; at least not soon. Therefore we have to consider their different approaches to aggregation and enhancement when assaying the relevance of projects like JSTOR, Hein-On-Line, or *LLMC-Digital* in a putative post-Google world. Here are just a few of the factors worth reflecting on.

— Focus: Because *LLMC-Digital* and its cousins focus on law, they can aggregate and organize collections tuned to the interests and needs of lawyers and those other researchers with a narrow interest in legal literature. In the Google approach, gazillions of data bits on an immeasurably broad array of subjects will be swimming in one vast sea of undifferentiated data. In cyberspace the equivalent of the fog of war may well be a bog of affinity. Even assuming that Google's trawling search engines can successfully navigate this ocean, will serious legal researchers really want to be competing for relevance rankings with, say, legions of e-Bayers? **(endnote 3)**

—Specialization: Not all search tools are equally efficient. Imagine an example from the analog world. Most people would prefer to have a one-volume *Complete Anthology of Early-American Poetry* to using the UofM's on-line catalog to search through all their 7-million volumes for "poetry, US, pre-1850." Of necessity, Google's search systems are basically one-size-fits-all. In contrast, sites like *LLMC-Digital* primarily serve one class of patron. They can afford to rifle-shoot available development funds toward system enhancements which best serve their own body of users. As important as having the needed resources, they are free to implement these focused enhancements without fearing flak from less sophisticated users. Our experience with *LLMC-Digital* provides an object lesson. We were able to "go up" as fast as we did because we piggybacked on a system (in some ways like Google's) developed previously by the UofM. However, the UofM search system was designed for a body of literature made up mostly of single-volume, non-legal treatises. It has taken us a year to make even modest modifications in that system to mold it to our needs, and we still have a fair

way to go. **(endnote 4)**

Page 3 begins here

— Timeliness: Even given the best of luck, Google's plans will take an uncertain number of years to implement. In the interim, the process will be so mechanized that pre-organization will be impossible. For example, at the UofM, the first swath of scanning will target the 3-million volumes housed in their Buhr remote storage facility. Those materials were "deselected-to-remote" precisely because no one had shown enough interest to check them out for years. Furthermore, they are shelved in random order by size, and that is the order in which they will be scanned. This doesn't mean that the world won't be glad to have them available. After all, these are those very "orphans" alluded to above. But it will be a long time before a sufficient critical mass builds up in any one subject area to justify the work of organizing it. Our patrons will not wait that long. Nor should we defer the potential side benefits realizable from our own digitization, not least massive space-recovery, to some uncertain future.

— Control: Our patrons are used to having their needs met by legal publishers and their law librarians. They would not take kindly to a course where we relied upon the vagaries of Google's "mining" of these libraries to eventually accumulate a usable digital collection of law books. Furthermore, they have always counted on us law librarians to acquire, organize, store and preserve the legal literature. Were we to rely upon the Google initiative to build the law library of the future, our control in all of these areas would be at least, and probably much more than, "once-removed."

— Bibliographic enhancement: Where we control the process from start to finish we can determine the level of extra effort and expense devoted to improving the bibliographic component of our digital collection. With *LLMC-Digital* for example, we have determined that nothing will go up on our site unless it has been fully cataloged to contemporary OCLC standards for digital titles. Does Google contemplate doing something similar for its 15-million-plus volumes? One needs to be dubious. In addition to providing for quality cataloging, both in the fiche era and now with digital, LLMC has always

recognized that some problem titles are almost unusable in their hardcopy guise, and would become worse with reformatting. In a fair number of instances **(endnote 5)** LLMC has provided guides or other aids which make the troublesome materials even more usable in their reformatted state than they were in the original. This sort of micro-care will be necessarily absent in mass-mining projects such as Google's. Of course, there will be nothing to stop folks from adding these features later, but we do it upfront.

Page 4 begins here

— Preservation: One of the primary duties of each generation of law librarians is to preserve the integrity of the intellectual content of those materials given to its care. Yet at few times is preservation more neglected than during the euphoria surrounding reformatting. For a majority of libraries, digital replacements will be primarily an opportunity to reclaim the high-quality space occupied by the original books. Too often the latter, either go straight to the dumpster, or glide there less dramatically, if just as mindlessly, after some years of conscience-lulling storage in ill-suited back rooms or leaky-piped basements. Meanwhile, few librarians know or seek to find out whether the reformatted replacements are complete as to pagination or in other respects. Relying for true preservation on a mass-mining project like Google's, or any other agency operating beyond our supervision, to attend to full preservation just won't cut it. The only way for us to be certain that the new generation of media is truly preserving its predecessor is for us to have a hand in its creation. Which brings us smoothly enough to….

De-accessioning As An Art

Thousands of law books were mindlessly discarded during the microfilm/microfiche era. This is not so much a criticism as a plain statement of fact. Space was tight. The books weren't used that much. And the replacement fiche were there. So hundreds of libraries tossed their paper sets of filmed materials, counting on "the other libraries" to retain preservation copies. Fortunately, not too much real damage was done. Fortuitously, using microforms was not popular with users. Many libraries decided to hold on to their hardcopy as long as possible. We therefore find ourselves today in a happy situation where large numbers of libraries still retain paper copies of materials which are now going up on the web.

During this round of reformatting, however, we think that the hardcopy-discard phenomenon will accelerate. **(endnote 6)** The digital copy on their desktop will be a lot more congenial to users than the fiche ever could be. Just as with the fiche, hardcopy will become backup. Since only so many paper copies of a given title will be needed for backup nationwide, many librarians will be tempted to reclaim valuable space. Deans will demand their "digital dividend," and, where possible, prudent librarians will give it to them; as they should.

Even so, once again it may just work out that no real damage is done. After all, we can all expect that at least some libraries (e.g. Harvard, Columbia, Michigan, Yale, etc.) will be holding on to their hardcopy for some titles for many years to come. We could take the laissez-faire approach followed in the fiche era and let every library just willy-nilly toss the books they don't want,

relying on others to keep them and hoping for the best.

However, this time around the omens are less auspicious. In the first place, the extent of discard is likely to be far greater. Secondly, the premise that "the big guys have everything" is more open to question. If we are to achieve the retention of some minimally-acceptable number of paper copies for all titles, we can't just presume and hope for the best. The only responsible way to achieve that goal is to take verifiable control of the national de-accession process. This means establishing responsibility coupled with accountability.

— Establishing responsibility requires identifying the likely ultimate repositories and obtaining reliable institutional commitments that they will take on that role.

— Achieving accountability requires that recordkeeping mechanisms be put in place to insure that the titles being retained by the depositories are complete, at least down to the volume level, and preferably down to the page.

Preservation of hardcopy is not a primary mission for LLMC. It is, however, something in which LLMC has a great interest. Our motivation for concern is our hands-on knowledge that this generation of digital-capture equipment is far from perfect, and our expectation that successor generations will have much improved capacities. There is a realistic likelihood that within this century the time will come when we will want to recapture

Page 5 begins here

from the original hardcopy and reformat much of the data now on our web sites. We want the hardcopy to be there if it is needed. So we are motivated to contribute where possible to the preservation process.

During the fiche era, and thereafter in our digital incarnation, we have been tracking the materials we filmed and scanned down to the image level. We keep track of missing pages and don't rest content until we have captured each page of every volume of every title we offer. The computer database to facilitate those goals for our purposes is already in place. With relatively minor effort and expense it could be modified to track the retention of a reasonable number of paper copies of those same titles down to the page/image level. It could also be expanded, if desired, to include non-LLMC titles. Finally, it could be mounted on the web so that libraries weeding a title could consult the database to determine whether gaps existed. Where gaps were found, the weeding library could offer its hardcopy to fill them. Where preservation targets already had been achieved, they could discard with a good conscience.

LLMC stands ready to cooperate with any responsible group of law libraries interested in implementing a national, post-digital, hard-copy preservation

program along the lines outlined above. (endnote 7) We would see our role in such a partnership as sharing the use of our database, paying for its modification to meet the hardcopy side of the program, and, if desired, managing the server so that the data-base could be accessed by interested libraries.

Report of the Interface Task Force

As reported earlier, at its July meeting the LLMC Board of Directors decided to establish a task force of librarians to provide user feedback regarding the priorities we should be setting in the use of our funds for improving the site. (endnote 8) Each of the Directors recruited members from their staffs or others known to them who had expertise using law-oriented digital services. A core group of ten techie and reference types was assembled under the chairmanship of Warren Rees of Notre Dame University Law Library.

The first report of the Task Force has now been received. It lists in order of priority the top ten improvements the group would like to see in the site. This list will be studied and further refined by the LLMC Board of Directors at its upcoming meeting in San Francisco on Jan. 8. It will then be referred to LLMC staff to work on, either in-house or, where appropriate, with our partners at the University of Michigan. The list is also provided here for the general membership with two goals. One, if something that has concerned you is mentioned here, then you have the assurance that it has reached the attention of the Board. Two, if the Task Force has not identified something that bugs you, then we need to hear from you now.

Report of the *LLMC-Digital* Interface Task force, with items listed in perceived priority:

— Searching options in LLMC Digital are confusing. When going in through the "Short Title List" or "Contents Status Table" it is not clear what you can search. Offering a link off each page with a description of the search options would be helpful. Also, it would be nice if an option were available to select specific volumes to search. As far as we could tell, one searches either volume by volume or in all volumes.

— Concern about the quality of the images on *LLMC-Digital* ranked very high. We don't know how much can be done about this, but we thought we should mention it if there is anything that can be done to improve the quality. (endnote 9)

Page 6 begins here

— Include a general description link from the home page that explains exactly what material is on the web site. This should be broken into specific topics to make it easier to select useful information.

—Use tables to structure the content of all the pages for a more consistent

look. (endnote 10) Also, some suggest using a sans serif font on every page.

— Show a path line that indicates where a user has been to get to the page currently on the screen and provide the option of jumping back to an earlier page by clicking on the path.

— Include a site map page to improve navigation.

— Use page templates that will make all pages appear uniform. Cascading style sheets were suggested by several.

— Include "contact us" links on all pages.

— Provide link to the free Adobe Acrobat reader from your pages.

— Break up long pages into multiple screens or at least provide navigation buttons frequently within the pages to make it easy to get back to the beginning.

In conclusion, on behalf of all the subscribers to *LLMC-Digital,* we would like to thank the Task Force members for the time, thought and energy they have put into this preliminary review. Working out solutions to some of the problems mentioned will be relatively easy. Others will take more time. But the first step along the journey has been taken and a precedent for constructively channeling user input has been established. For that we are all grateful

A Friend in Need

Most subscribers to LLMC may not know that over the years we have had three local Hawaii sources for perhaps a third of our titles. The Hawaii Supreme Court Library provided much of the hardcopy for our state court collections and our collections of selected cases and legal encyclopedias. Later on, as the University of Hawaii's law library grew, we usefully borrowed much of their law review stock and the pre-copyright NRS. Our primary local source for U.S. federal documents and international organization material was the GovDocs Collection of Hamilton Library, the main university library building.

Last Halloween Day disaster hit Hamilton. During exceptionally heavy rains, debris washed down a small river called Manoa Stream accumulated at a bridge, about a mile upland from the Manoa Campus. The bridge became a dam, diverting floodwaters into the local city streets, down which they flowed onto the campus. Hamilton Library has a full basement, in which were housed both the GovDocs Collection and the Maps Collection. The basement being the functional equivalent of a well, floodwaters filled that well up to about the six-foot level. Both collections were mostly destroyed. Over 2-million items were lost. Total damage on campus is now estimated to exceed $100-million. (endnote 11)

On behalf of all of the LLMC libraries who have benefited from our

partnership with them, LLMC extends its deep condolences to our good friends at Hamilton GovDocs. We have offered to do whatever we reasonably can to help them rebuild their collections.

Endnotes:

1. Actually, we can imagine it. The likely work engine would be the Kirtas APT BookScan 1200 (for a video demo see www.kirtas-tech.com). This robotic digital book scanner boasts a capture rate of 1,200 pages per hour (pph). Since no data-capture machinery ever achieves its claimed top throughput, it might be safer to use a figure of about 1,000 pph. Hitting ASCII's roughly 20,000 pph would thus require a bank of about 20 Kirtas machines, with each machine scanning roughly 6.6 books per hr. At a list price of $150K per machine that comes to only $3-million or less in capital costs, with Google's likely volume discount probably balanced out by the need to buy backup units.

Global labor costs are also within the realm of estimate. At about 8 minutes per book, one FTE operator probably could keep one Kirtas machines supplied with books for an 8-hr shift, Frontline operators usually needs backup by at least another half FTE person working elsewhere in the production process. That translates to ca. 30 FTE per shift and ca. 90 FTE per 24/7 day. Figuring an average pay scale of $17 per hr (wages+benefits), labor costs would be about $12,240 per day, or roughly $4.47-million in the first year. Top that off with about 30% overhead for supervision and management to arrive at first-year labor costs of about $6.8-million. Factor in an average inflation of 3% over the 6-year life of the project and total labor costs come to something like $44-million,

Applying a generous contingency allowance of about 10% to the total (capital costs+labor) of ca. $47-million, one arrives at a rough, but usable, estimate of perhaps $51.7-million for the on-site data-capture phase of the UofM portion of the project. Our experience with *LLMC-Digital* tells us that post-production processing-of-images costs can add as much as 40% to data-capture expenditures. So our back-of-an-envelope estimate of the cost to Google for doing UofM's 7-million volumes totals to something like $72.4-million. This seems to validate Google's own estimate (reported in the New York Times) of global costs at about $10.00 per volume.

$70-million give-or-take isn't chump change, but it's not likely to exhaust the discretionary cash to be found in Google's deep pockets.

2. Some of the following reflections were contributed by Maria Bonn,

principal contact at our LLMC-Digital partner, the UofM Scholarly Publishing Office and Margaret Leary of that university's law library.

3. To say nothing of the potential extra static caused by Google's practice of allowing interested publishers to pay to boost their rankings. Of course, it's not likely that we would soon see Podunk Law School subsiding more hits on its law review over, say, sweet irony, that of the UofM. But what of the potential static from publishers totally outside the law field?

4. See e.g., the discussion of the first report of the *LLMC-Digital* Interface Task Force above, p.4.

5. One example would be the *Harvard Annual legal Bibliography* (HALB), which, by the time LLMC took it on for filming, consisted of 21 annuals, each of which one had to consult to perform a simple search. A difficult enough problem when using the books, it would have been a nightmare when juggling fiche. Realizing that it would be worse than useless to film the original books, LLMC obtained access from Harvard to the original 575,000 catalog cards from which the HALB books had been photo printed. All of these entries were reorganized into one efficient research tool, supplemented with additional entries which had been missed in the original printings, and enhanced with timesaving tables. In the guise of a microfiche edition, the HALB was reborn as a streamlined reference tool, reducing search time within those twenty years of Harvard's periodical holdings from an hour or more to minutes.

Another example would be *The Persian Gulf Gazette,* which was recently digitally scanned and will be available on *LLMC-Digital* early in 2005. Again, the hardcopy was virtually unusable. Texts of orders, rules, notices, proposed legislation, etc. are scattered randomly throughout the 20 years of its publication. There is no index. Finding something meant browsing through 20 volumes; a daunting project with the books, an impossible prospect with fiche, and likely to be a laborious trudge even with digital. LLMC put in the time and money to create a 16-page "Table of Contents by Jurisdiction," which will be joined with the title on-line. The table segregates materials for all eight jurisdictions covered by the Gazette, with entries organized chronologically by date of publication therein. No doubt this is a luxury, but it is a luxury we can afford because we have a specialized clientele, whom we know will appreciate and

use it.

6. For a fuller discussion on this topic see this newsletter, Issue #9, p. 4, column 2.

7. Along these lines it is noteworthy that a committee set up by the LIPA group is having a "brain-storming session" on the general topic of post-digital hardcopy preservation during the upcoming AALS convention in San Francisco from 10:30–noon on Saturday, Jan. 8. Because the meeting space has a limited capacity, persons interested in participating in this session should e-mail the Chair, Kent McKeever of Columbia, before Jan. 6. His address is mckeever@la.columbia.edu.

<>

8. The task force is discussed in previous Newsletters at: Issue # 9, p.4 & Issue # 10, p. 2..

9. Readers should bear in mind that this problem of image quality has two distinct faces: a.) establishing a proper balance between images scanned from our fiche backfile and images scanned *de novo* from original hardcopy, and b.) the problem of poor quality paper originals which will neither film nor scan at acceptable levels. The LLMC staff and Directors have been wrestling with these questions for some time now, and a definite statement of policy is likely to be announced soon and explained in the Newsletter early in the new year.

10. To avoid possible confusion, it is worth noting that when the Task Force alludes to restructuring or reformatting "pages" (items 4, 7 & 10), this means only those pages of text developed in-house as explanatory or site-navigation material by LLMC staff or our partners at Michigan. There is no intent that LLMC would ever violate its core goal of providing exact copies of the pages in the original legal material we provide on the site..

11. A sequence of flood images can be viewed at http://libweb.hawaii.edu/uhmlib/news/flood-articles.html. Articles and images also are avail-able at http://www.kaleo.org/vnews/display.v/ART/2004/11/01/4185e690886da. Finally, a melancholic view of the soggy remains of the Gov-Docs copy of the *US Statutes at Large,* the original for both our fiche and the copy now on *LLMC-Digital,* is available at http://www.drdriving.org/flood/october_flood-Pages/Image84.html.